

# Introduction to Machine Learning

Journées Calcul et Apprentissage

---

Aurélien Garivier

24-25 avril 2019



# Table of contents

1. What is Machine Learning?
  - Data and Learning Algorithms
  - Classification Framework
2. First Algorithms: fitting versus generalizing
3. Nearest-Neighbor Classification
4. Empirical Risk Minimization
5. Support Vector Machines
6. Neural Networks

# What is Machine Learning?

---

# Why Machine Learning?

Actualité

## Yann LeCun, Geoffrey Hinton et Yoshua Bengio reçoivent le prix Turing

Par Stéphane Nachez - 27 mars 2019



### LE MACHINE LEARNING PROVOQUE UNE CRISE DANS LE DOMAINE DE LA SCIENCE

Basilien L. - 19 février 2019 - Analytics, Data Analytics, Intelligence artificielle - 1 commentaire

*Le Machine Learning est en train de provoquer une grave crise de reproductibilité dans le domaine de la science. C'est ce qu'affirme la statisticienne Genevera Allen de la Rice University dans le cadre de la conférence AAAS Annual Meeting.*

De plus en plus de chercheurs utilisent le **Machine Learning** pour analyser des données et y détecter des tendances. Cependant, dans le cadre de la conférence scientifique AAAS Annual Meeting, la statisticienne Genevera Allen de la Rice University a tenu à tirer la sonnette d'alarme. Selon elle, le Machine Learning est en passe de provoquer **une crise de reproductibilité dans le domaine de la science.**

SHARE SPECIAL VIEWPOINTS



## Machine Learning for Science: State of the Art and Future Prospects

Eric Mjølhusnes<sup>1</sup>, Dennis DeCoste

<sup>1</sup> See all authors and affiliations

Science, 14 Sep 2021.  
Vol. 293, Issue 5537, pp. 2031-2035.  
DOI: 10.1126/science.200.5507.2031

Article

Figures & Data

Info & Metrics

eLetters

PDF

### Abstract

Recent advances in machine learning methods, along with successful applications across a wide variety of fields such as planetary science and bioinformatics, promise powerful new tools for practicing scientists. This viewpoint highlights some useful characteristics of modern machine learning methods and their relevance to scientific applications. We conclude with some speculations on near-term progress and promising directions.

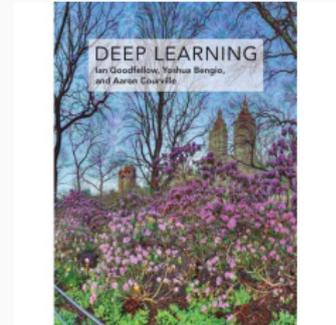
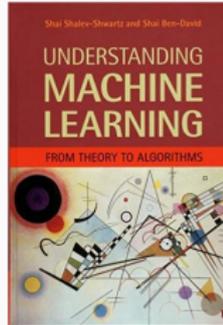
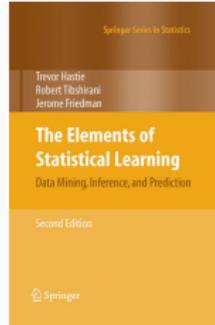
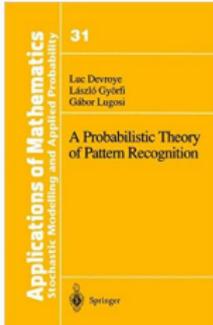
PUBLIC RELEASE: 15-FEB-2019

## Can we trust scientific discoveries made using machine learning?

*Rice U. expert: Key is creating ML systems that question their own predictions*

RICE UNIVERSITY

# Where to learn more?



# What do I need to practice ML? R

RStudio interface showing R code and a plot.

```
R> # Source on Save | Source | Run | Stop | Source |
254 plot(TT, RR[,j], type="l", lwd=4, lty=lty[j], log="x", xlim=c(min(TT), max(TT)), ylim = c(0, max(RR[
255 #arrow(TT, RR[,j+2*SD[1,]/sqrt(Nmc), TT, RR[,j+2*SD[1,]/sqrt(Nmc), length=0.1, lwd=4, code=3, lo
256 # for (j in 1:n0Strategies){
257 # lines(TT, RR[,j], lwd=4, lty=lty[j])
258 # r <- |a|RR[,j]-TT|*10000 - log(TT/|TT-10000|))
259 # lines(TT/500, r, col="red", lty="n", col="red", lty=lty[j])
260 # slopes[j] <- rcoefficient[2]/4
261 #arrow(TT, RR[,j+2*SD[1,]/sqrt(Nmc), TT, RR[,j+2*SD[1,]/sqrt(Nmc), length=0.1, lwd=4, code=3, lo
262 # }
263 #for (h in c(0.5, 1, 2, 4)){ lines(TT, h*log(TT^0.2/h)/d, col="red", lwd=2) }
264
265 #plot(TT, dRR[,j]/log(TT), type="l", lwd=4, lty=lty[j], log="x", xlim=c(min(TT), max(TT)), ylim = c
266 #arrow(TT, RR[,j+2*SD[1,]/sqrt(Nmc), TT, RR[,j+2*SD[1,]/sqrt(Nmc), length=0.1, lwd=4, code=3, lo
267 #for (j in 2:n0Strategies){
268 # lines(TT, dRR[,j]/log(TT), lwd=4, lty=lty[j])
269 #arrow(TT, RR[,j+2*SD[1,]/sqrt(Nmc), TT, RR[,j+2*SD[1,]/sqrt(Nmc), length=0.1, lwd=4, code=3, lo
270 # }
271 #names <- c("FB-ETC", "SPRT-ETC", "BAI-ETC", "D-UCB", "UCB") #??-UCB, "UCB")
272
273 order <- c(1,3,5,2,4)
274 legend(50, 80, as.list(order, function(k) paste(names[k], ":", round(slopes[k], 2))), lty = lty[order
275
276
277
278 # T <- c(2.5, 10, 20, 30, 50, 100, 150, 200, 300, 400, 500, 600, 800, 1000, 1500, 2000, 3000, 5000, 8000, 10000)/d/2; #T;
279 # T <- c(2.5, 10, 20, 30, 50, 100, 150, 200, 300, 400, 500, 600, 800, 1000)/d/2; #mc <- 10000; createData(T, #T);
280 # T <- c(10, 20)/d/2; #mc <- 1000; createData(F, #mc);
281 # T <- c(100, 800, 1500)/d/2; #mc <- 10000; createData(T, #mc);
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
267
```

# What do I need to practice ML? python

Activités mer, 19:17 Spyder (Python 2.7)

Fichier Édition Recherche Source Exécution Débugger Consoles Outils Affichage Aide

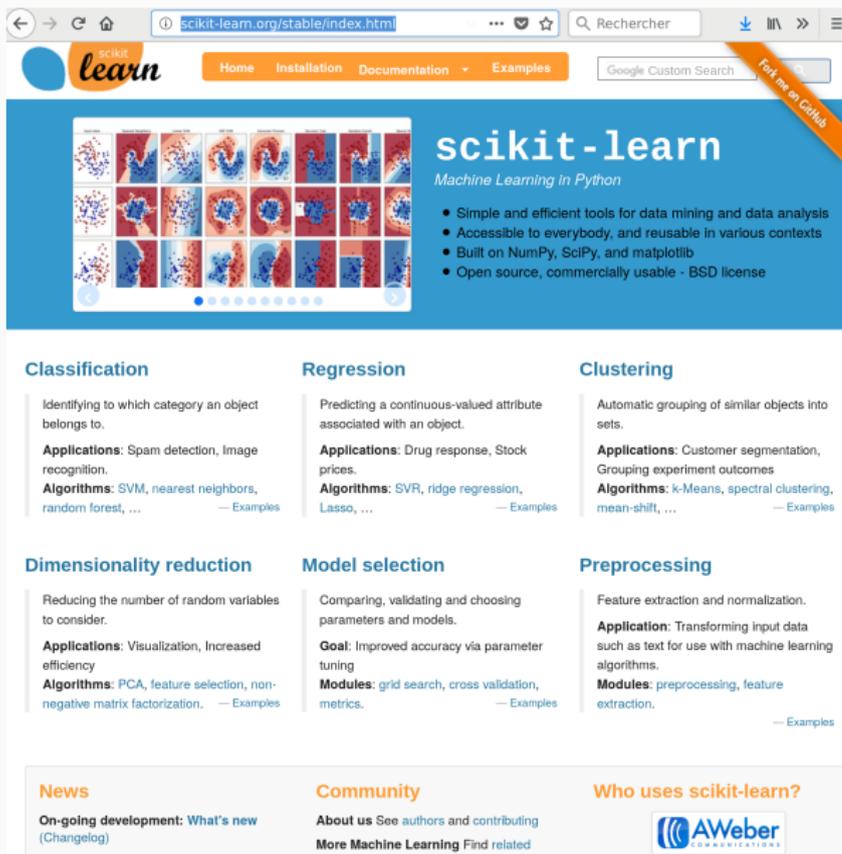
Editeur - /home/agarvis/ownCloud/prog/python/HC/deprecated/kaplan\_meier.py

```
1 #-*- coding: utf-8 -*-
2 Created on Tue Aug 16 03:05:03 2016
3
4 @author: agarvis
5
6
7
8 import numpy as np
9 from random import randint
10 import matplotlib.pyplot as plt
11
12
13 def kaplan_meier(L, Delta, censorData):
14     N = len(Delta)-1
15     atrisk = np.zeros(N) # alive after Delta[i]
16     survived = np.zeros(N) # died between Delta[i] and Delta[i+1]
17     for x in L:
18         while 1 <= N and x[0] <= Delta[i] <= censorData and x[0] <= Delta[i] + x[1]:
19             atrisk[i] += 1
20             if x[0] <= Delta[i] <= x[1]:
21                 survived[i] += 1
22             1 += 1
23     S = np.concatenate(([], np.cumprod(survived/atrisk)))
24     return(S)
25 # arbitrary: index shift of 1? see kaplan_meier_2 which conforms to package
26
27 def plot_coxall():
28     plot_coxall()
29
30 n = 10000
31 x = [np.casus(randint(0, random()) for k in range(n))]
32 time = 0.7
33 k = [int(i), int(i+1)]*new*(2-int(i)) for z in x]
34
35 Delta = np.array([float(i)/N for i in range(N+1)])
36
37 S = kaplan_meier(x, Delta, new)
38 print(S)
39
40 # plt.subplot(2, 1, 1)
41 # plt.hold('on')
42 # plt.plot(Delta, S, Delta)
43 # plt.hold('off')
44 # plt.xlabel('0.05, 2.0, 11')
45
46
47 def kn_nact(E):
48     I = np.arange(0)
49     t = [t[i] for k in I]
50     q = [E[i] for k in I]
51     n = len(I)
52     S = np.ones(n)
53     H = np.zeros(n)
54     for k in range(0, 1):
55         S[k+1] = S[k] * (n-k+1)/(n-k)
56         H[k+1] = H[k] + E[k]/(n-k+1)
57
```

Console (Python)

```
Python 2.7.10 [Anaconda 2.3.0 (64-bit)] (default, Sep 15 2015, 14:50:01)
Type "copyright", "credits" or "license()" for more information.
>
Python 4.0.0 - An enhanced Interactive Python.
? -> Introduction and overview of Python's features.
help() -> Quick reference.
help() -> Python's own help system.
object? -> Details about 'object' use 'object??' for extra details.
__main__? -> A brief reference about the graphical user interface.
In [1]: runfile('/home/agarvis/ownCloud/prog/python/HC/payment_renewal_study_myasoothazard.py',
wdir='/home/agarvis/ownCloud/prog/python/HC')
Traceback (most recent call last):
  File "c:\python-input-1-9463842892", line 1, in <module>
    runfile('/home/agarvis/ownCloud/prog/python/HC/payment_renewal_study_myasoothazard.py',
wdir='/home/agarvis/ownCloud/prog/python/HC')
  File "/home/agarvis/anaconda/lib/python2.7/site-packages/ipython/lib/widgets/externalshell/sitecustomize.py", line 685, in runfile
    execute(filename, namespace)
  File "/home/agarvis/anaconda/lib/python2.7/site-packages/ipython/lib/widgets/externalshell/sitecustomize.py", line 78, in execute
    builtins.execute(filename, *where)
  File "/home/agarvis/ownCloud/prog/python/HC/payment_renewal_study_myasoothazard.py", line 60, in main
    (t, S) = kn(E)
NameError: name 'kn' is not defined
In [2]: runfile('/home/agarvis/ownCloud/prog/python/HC/deprecated/kaplan_meier.py',
wdir='/home/agarvis/ownCloud/prog/python/HC/deprecated')
In [3]:
```

Figure 2



scikit-learn

Home Installation Documentation Examples

Google Custom Search

Fork me on GitHub

## scikit-learn

Machine Learning In Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

### Classification

Identifying to which category an object belongs to.

**Applications:** Spam detection, Image recognition.

**Algorithms:** SVM, nearest neighbors, random forest, ... — Examples

### Regression

Predicting a continuous-valued attribute associated with an object.

**Applications:** Drug response, Stock prices.

**Algorithms:** SVR, ridge regression, Lasso, ... — Examples

### Clustering

Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, Grouping experiment outcomes

**Algorithms:** k-Means, spectral clustering, mean-shift, ... — Examples

### Dimensionality reduction

Reducing the number of random variables to consider.

**Applications:** Visualization, Increased efficiency

**Algorithms:** PCA, feature selection, non-negative matrix factorization. — Examples

### Model selection

Comparing, validating and choosing parameters and models.

**Goal:** Improved accuracy via parameter tuning

**Modules:** grid search, cross validation, metrics. — Examples

### Preprocessing

Feature extraction and normalization.

**Application:** Transforming input data such as text for use with machine learning algorithms.

**Modules:** preprocessing, feature extraction. — Examples

### News

On-going development: [What's new \(Changelog\)](#)

### Community

About us See [authors](#) and [contributing](#)

[More Machine Learning](#) Find related

### Who uses scikit-learn?



What is Machine Learning?

Data and Learning Algorithms

Classification Framework

First Algorithms: fitting versus generalizing

Nearest-Neighbor Classification

Empirical Risk Minimization

Support Vector Machines

Neural Networks

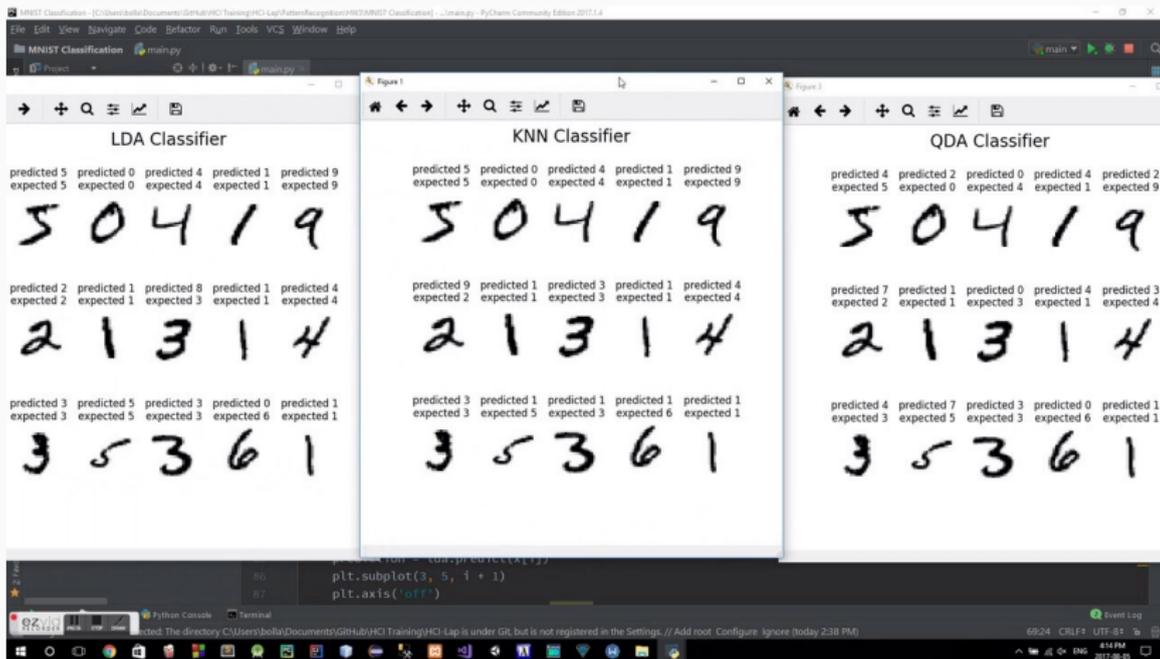
# What is Machine Learning?

- Algorithms operate by building a model from **example** inputs in order to make data-driven **predictions or decisions**...
- ...rather than following strictly static program instructions: useful when designing and programming explicit algorithms is unfeasible or poorly efficient.

## Within Artificial Intelligence

- evolved from the study of pattern recognition and computational learning theory in artificial intelligence.
- AI: emulate cognitive capabilities of humans (big data: humans learn from abundant and diverse sources of data).
- a machine mimics "cognitive" functions that humans associate with other human minds, such as "learning" and "problem solving".

# Example: MNIST dataset



# Definition

## **Arthur Samuel (1959)**

Field of study that gives computers the ability to learn without being explicitly programmed

## **Tom M. Mitchell (1997)**

A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$  if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ .

# Machine Learning: Typical Problems

- spam filtering, text classification
- optical character recognition (OCR)
- search engines
- recommendation platforms
- speech recognition software
- computer vision
- bio-informatics, DNA analysis, medicine
- ...

For each of this task, it is possible but very inefficient to write an explicit program reaching the prescribed goal.

It proves much more succesful to have a machine infer what the good decision rules are.

# What is Statistical Learning?

- = Machine Learning using statistics-inspired tools and guarantees
  - Importance of **probability**- and **statistics**-based methods  
→ **Data Science** (Michael Jordan)
  - **Computational Statistics**: focuses in prediction-making through the use of computers together with statistical models (ex: Bayesian methods).
  - **Data Mining** (unsupervised learning) focuses more on exploratory data analysis: discovery of (previously) unknown properties in the data. This is the analysis step of Knowledge Discovery in Databases.
  - Machine Learning has more **operational** goals  
Ex: ~~consistency~~ → oracle inequalities  
Models (if any) are *instrumental*.  
ML more focused on *correlation*, less on *causality* (now changing).
  - Strong ties to **Mathematical Optimization**, which furnishes methods, theory and application domains to the field

# What is ML composed of?



What is Machine Learning?

Data and Learning Algorithms

Classification Framework

First Algorithms: fitting versus generalizing

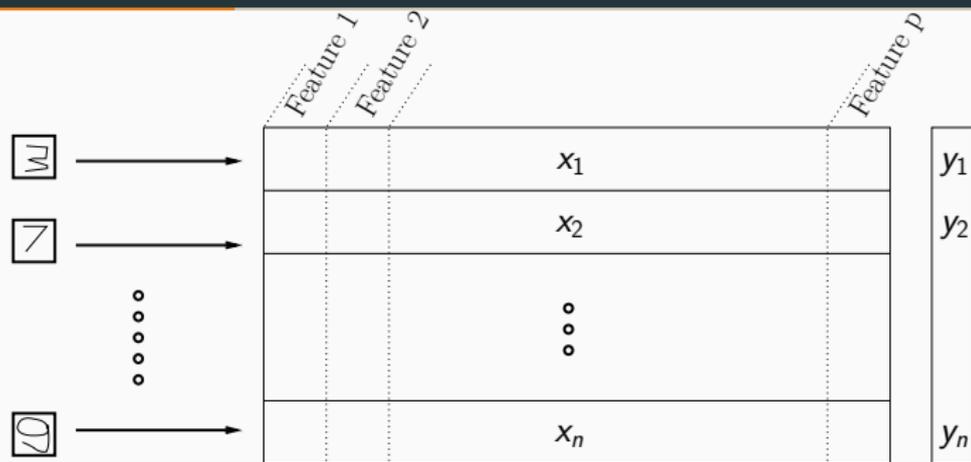
Nearest-Neighbor Classification

Empirical Risk Minimization

Support Vector Machines

Neural Networks

# What is a classifier?



$$X \in \mathcal{M}_{n,p}(\mathbb{R})$$

$$Y \in \mathcal{Y}^n$$

Data:  $n$ -by- $p$  matrix  $X$

- $n$  examples = points of observations
- $p$  features = characteristics measured for each example

Classifier  $\mathcal{A}_n$

$$\begin{array}{c} \downarrow \\ h_n : \mathcal{X} \rightarrow \mathcal{Y} \\ \boxed{6} \mapsto 6 \end{array}$$

- Inside R: package datasets
- Inside python/scikitlearn: package `sklearn.datasets`
- UCI Machine Learning Repository



- Challenges: Kaggle, etc.

## What is Machine Learning?

Data and Learning Algorithms

Classification Framework

First Algorithms: fitting versus generalizing

Nearest-Neighbor Classification

Empirical Risk Minimization

Support Vector Machines

Neural Networks

# Statistical Learning Hypothesis

## Assumption

- The examples  $(X_i, Y_i)_{1 \leq i \leq n}$  are iid samples of an unknown joint distribution  $\mathcal{D}$ ;
- The points to classify later are also independent draws of the *same* distribution  $\mathcal{D}$ .

Hence, for every *decision rule*  $h : \mathcal{X} \rightarrow \mathcal{Y}$  we can define the *risk*

$$L_{\mathcal{D}}(h) = \mathbb{P}_{(X,Y) \sim \mathcal{D}}(h(X) \neq Y) = \mathcal{D}\left(\{(x,y) : h(x) \neq y\}\right).$$

The goal of the learning algorithm is to *minimize the expected risk*:

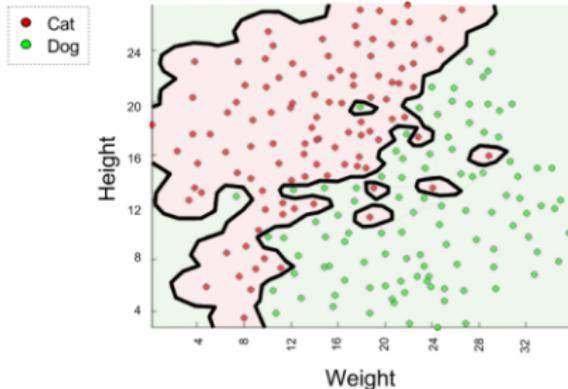
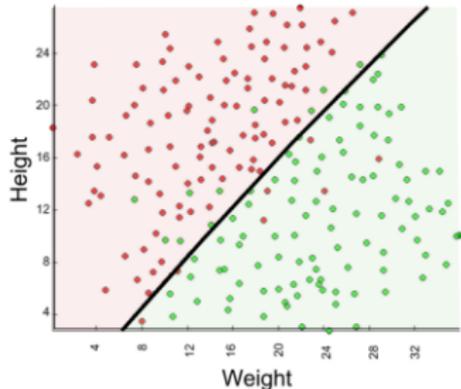
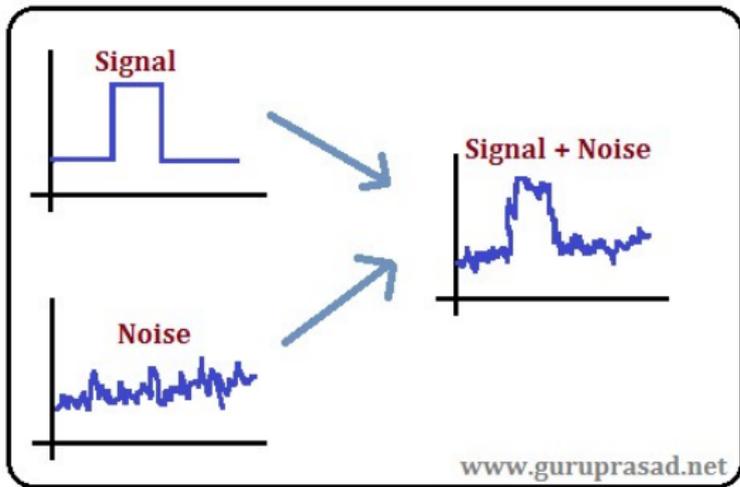
$$R_n(\mathcal{A}_n) = \mathbb{E}_{\mathcal{D}^{\otimes n}} \left[ L_{\mathcal{D}} \left( \underbrace{\mathcal{A}_n((X_1, Y_1), \dots, (X_n, Y_n))}_{\hat{h}_n} \right) \right]$$

for every distribution  $\mathcal{D}$ , using only the examples.

# Signal and Noise

*new york times bestseller*  
*noise and the noi*  
*the signal and th*  
*and the noise and*  
*the noise and th*  
*why so many noi*  
*predictions fail—*  
*but some don't th*  
*and the noise and*  
*nate silver the nc*

"Could turn out to be one of the more momentous books of the decade." —*The New York Times Book Review*



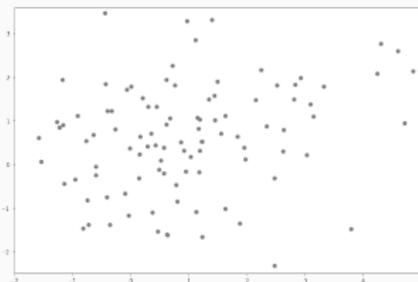
## Example: Character Recognition

Domain set $\mathcal{X}$	28 × 28 images
Label set $\mathcal{Y}$	{0, 1, ..., 9}
Joint distribution $\mathcal{D}$	?
Prediction function $h \in \mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$	
Risk $R(h) = P_{\mathcal{D}}(h(X) \neq Y)$	
Sample $S_n = \{(X_i, Y_i)\}_{i=1}^n$	MNIST dataset
Empirical risk $L_S(h) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{h(X_i) \neq Y_i\}$	
Learning algorithm $\mathcal{A} = (\mathcal{A}_n)_n, \mathcal{A}_n : (\mathcal{X} \times \mathcal{Y})^n \rightarrow \mathcal{H}$	neural nets, boosting...
Expected risk $R_n(\mathcal{A}) = \mathbb{E}_n [L_{\mathcal{D}}(\mathcal{A}_n(S_n))]$	

## Two visions of $\mathcal{D}$

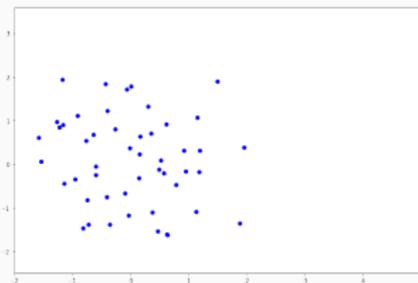
As a pair  $(\mathcal{D}_x, k)$ , where

- for  $A \subset \mathcal{X}$ ,  $\mathcal{D}_x(A) = \mathcal{D}(A \times \mathcal{Y})$  is the marginal distribution of  $X$ ,
- and for  $x \in \mathcal{X}$  and  $B \subset \mathcal{Y}$ ,  
 $k(B|x) = \mathcal{D}(Y \in B|X = x)$  is (a version of) the conditional distribution of  $Y$  given  $X$ .



As a pair  $(\mathcal{D}_y, (\mathcal{D}(\cdot|y))_y)$ , where

- for  $y \in \mathcal{Y}$ ,  $\mathcal{D}_y(y) = \mathcal{D}(\mathcal{X} \times y)$  is the marginal distribution of  $Y$ ,
- and for  $A \subset \mathcal{X}$  and  $y \in \mathcal{Y}$ ,  
 $\mathcal{D}(A|y) = \mathcal{D}(X \in A|Y = y)$  is the conditional distribution of  $X$  given  $Y = y$ .



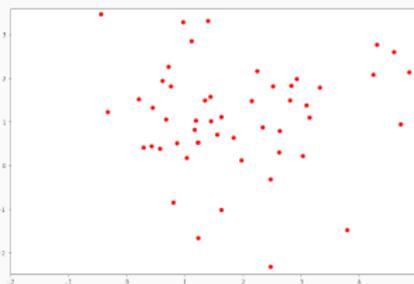
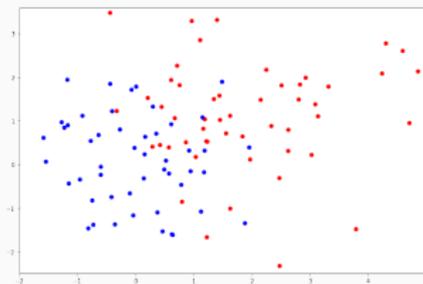
## Two visions of $\mathcal{D}$

As a pair  $(\mathcal{D}_x, k)$ , where

- for  $A \subset \mathcal{X}$ ,  $\mathcal{D}_x(A) = \mathcal{D}(A \times \mathcal{Y})$  is the marginal distribution of  $X$ ,
- and for  $x \in \mathcal{X}$  and  $B \subset \mathcal{Y}$ ,  
 $k(B|x) = \mathcal{D}(Y \in B|X = x)$  is (a version of) the conditional distribution of  $Y$  given  $X$ .

As a pair  $(\mathcal{D}_y, (\mathcal{D}(\cdot|y))_y)$ , where

- for  $y \in \mathcal{Y}$ ,  $\mathcal{D}_y(y) = \mathcal{D}(\mathcal{X} \times y)$  is the marginal distribution of  $Y$ ,
- and for  $A \subset \mathcal{X}$  and  $y \in \mathcal{Y}$ ,  
 $\mathcal{D}(A|y) = \mathcal{D}(X \in A|Y = y)$  is the conditional distribution of  $X$  given  $Y = y$ .



## Performance Limit: Bayes Classifier

Consider binary classification  $\mathcal{Y} = \{0, 1\}$ ,  $\eta(x) := \mathcal{D}(Y = 1|X = x)$ .

### Theorem

*The Bayes classifier is defined by*

$$h^*(x) = \mathbb{1}\{\eta(x) \geq 1/2\} = \mathbb{1}\{\eta(x) \geq 1 - \eta(x)\} = \mathbb{1}\{2\eta(x) - 1 \geq 0\}.$$

*For every classifier  $h : \mathcal{X} \rightarrow \mathcal{Y} = \{0, 1\}$ ,*

$$L_{\mathcal{D}}(h) \geq L_{\mathcal{D}}(h^*) = \mathbb{E}\left[\min(\eta(X), 1 - \eta(X))\right].$$

*The Bayes risk  $L_{\mathcal{D}}^* = L_{\mathcal{D}}(h^*)$  is called the **noise** of the problem.*

*More precisely,*

$$L_{\mathcal{D}}(h) - L_{\mathcal{D}}(h^*) = \mathbb{E}\left[|2\eta(X) - 1| \mathbb{1}\{h(X) \neq h^*(X)\}\right].$$

Extends to  $|\mathcal{Y}| > 2$ .

## **First Algorithms: fitting versus generalizing**

---

# Nearest-Neighbor Classification

---

# The Nearest-Neighbor Classifier

We assume that  $\mathcal{X}$  is a metric space with distance  $d$ .

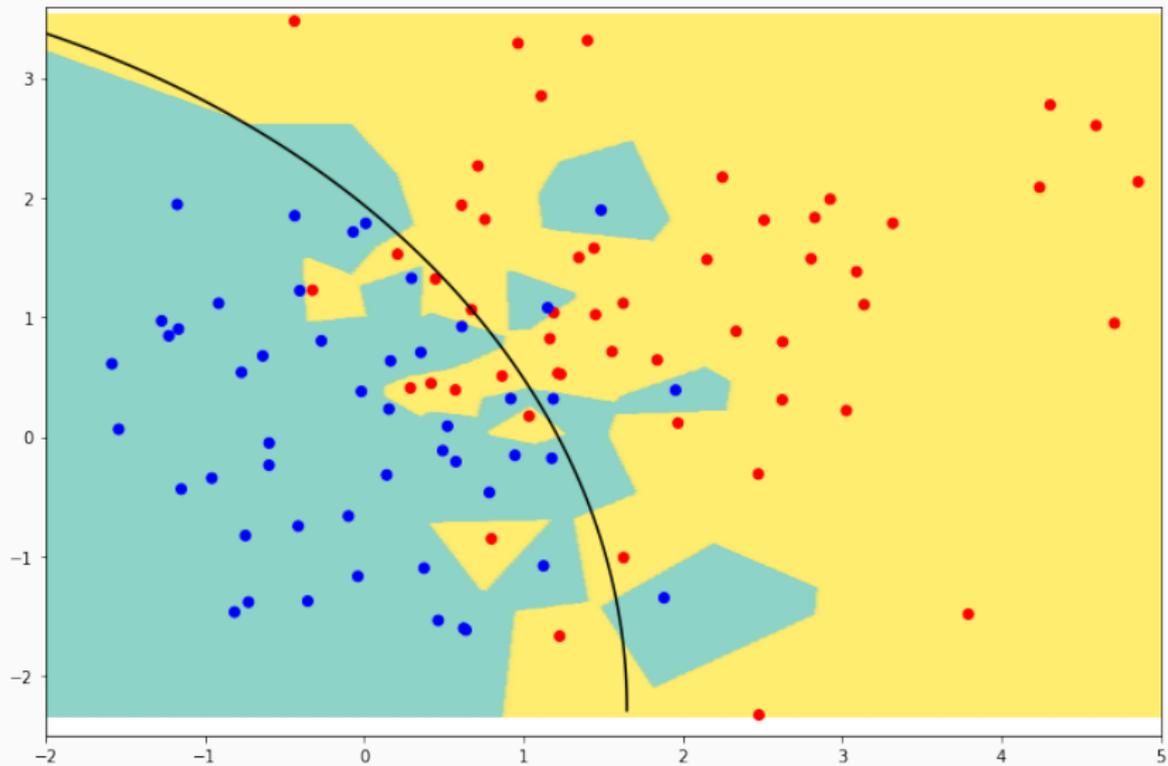
The nearest-neighbor classifier  $\hat{h}_n^{NN} : \mathcal{X} \rightarrow \mathcal{Y}$  is defined as

$$\hat{h}_n^{NN}(x) = Y_I \text{ where } I \in \arg \min_{1 \leq i \leq n} d(x - X_i).$$

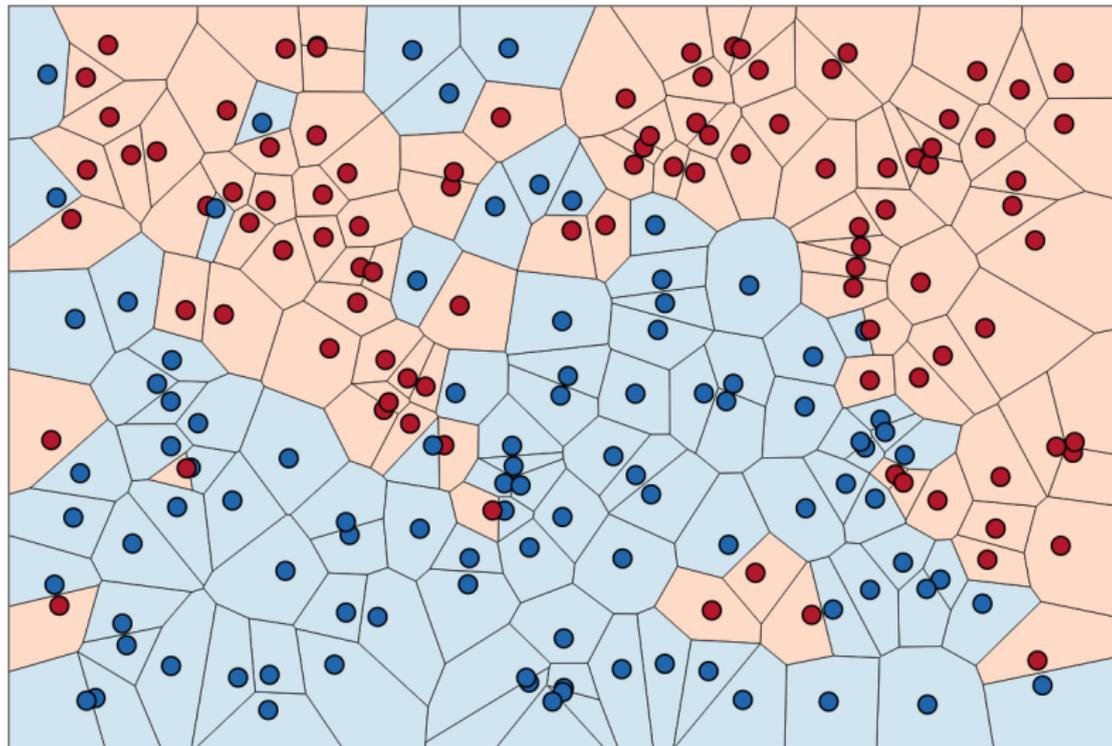
Typical distance:  $L^2$  norm on  $\mathbb{R}^d$ :  $\|x - x'\| = \sqrt{\sum_{j=1}^d (x_j - x'_j)^2}$ .

Buts many other possibilities: Hamming distance on  $\{0, 1\}^d$ , etc.

# Numerically



# Numerically



# The most simple analysis of the most simple algorithm

**A1.**  $\mathcal{Y} = \{0, 1\}$ .

**A2.**  $\mathcal{X} = [0, 1]^d$ .

**A3.**  $\eta$  is  $c$ -Lipschitz continuous:

$$\forall x, x' \in \mathcal{X}, |\eta(x) - \eta(x')| \leq c \|x - x'\| .$$

## Theorem

*Under the previous assumptions, for all distributions  $\mathcal{D}$  and all  $m \geq 1$*

$$L_{\mathcal{D}}(\hat{h}_n^{NN}) \leq 2L_{\mathcal{D}}^* + \frac{3c\sqrt{d}}{n^{1/(d+1)}} .$$

## Proof Outline

- Conditioning: as  $I(x) = \arg \min_{1 \leq i \leq n} \|x - X_i\|$ ,

$$L_D(\hat{h}_n^{NN}) = \mathbb{E} \left[ \mathbb{E} \left[ \mathbb{1}\{Y \neq Y_{I(X)}\} \mid X, X_1, \dots, X_n \right] \right].$$

- $Y \sim \mathcal{B}(p)$ ,  $Y' \sim \mathcal{B}(q) \implies \mathbb{P}(Y \neq Y') \leq 2 \min(p, 1-p) + |p - q|$ ,  
 $\mathbb{E} \left[ \mathbb{1}\{Y \neq Y_{I(X)}\} \mid X, X_1, \dots, X_n \right] \leq 2 \min(\eta(X), 1-\eta(X)) + c \|X - X_{I(X)}\|.$

- Partition  $\mathcal{X}$  into  $|\mathcal{C}| = T^d$  cells of diameter  $\sqrt{d}/T$ :

$$\mathcal{C} = \left\{ \left[ \frac{j_1 - 1}{T}, \frac{j_1}{T} \right] \times \dots \times \left[ \frac{j_d - 1}{T}, \frac{j_d}{T} \right], \quad 1 \leq j_1, \dots, j_d \leq T \right\}.$$

- 2 cases: either the cell of  $X$  is occupied by a sample point, or not:

$$\|X - X_{I(X)}\| \leq \sum_{c \in \mathcal{C}} \mathbb{1}\{X \in c\} \left( \frac{\sqrt{d}}{T} \mathbb{1} \bigcup_{i=1}^n \{X_i \in c\} + \sqrt{d} \mathbb{1} \bigcap_{i=1}^n \{X_i \notin c\} \right).$$

- $\implies \mathbb{E}[\|X - X_{I(X)}\|] \leq \frac{\sqrt{d}}{T} + \frac{\sqrt{d}T^d}{en}$  and choose  $T = \lfloor n^{\frac{1}{d+1}} \rfloor$ .

# What does the analysis say?

- Is it loose? (sanity check: uniform  $\mathcal{D}_X$ )
- *Non-asymptotic* (finite sample bound)
- The second term  $\frac{3c\sqrt{d}}{n^{1/(d+1)}}$  is *distribution independent*
- Does not give the trajectorial decrease of risk
- Exponential bound  $d$  (cannot be avoided...)  
 $\implies$  *curse of dimensionality*
- How to improve the classifier?

Let  $\mathcal{X}$  be a (pre-compact) metric space with distance  $d$ .

## $k$ -NN classifier

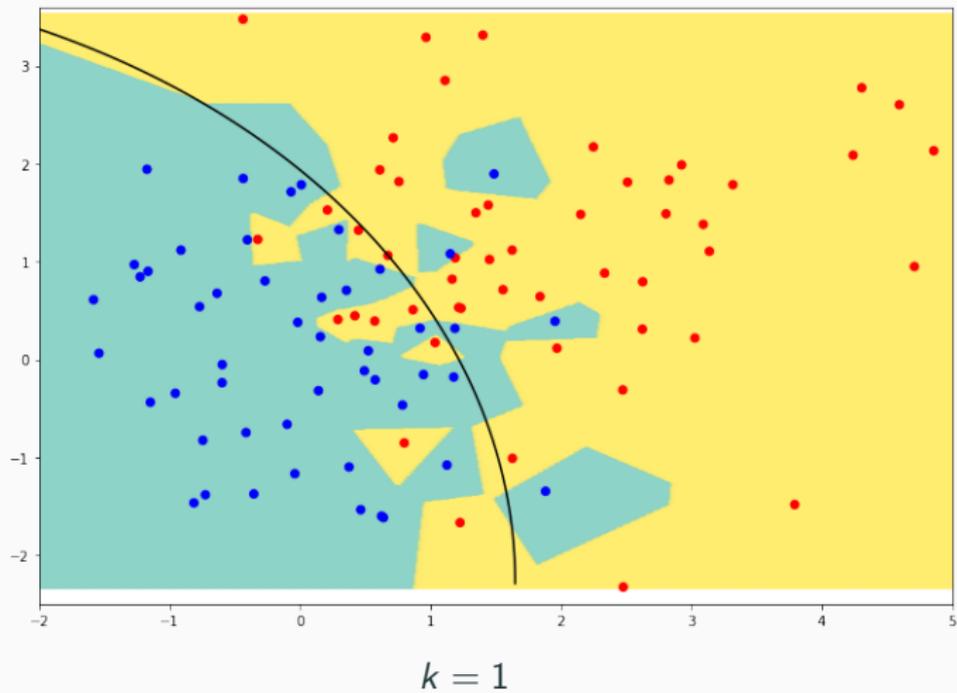
$h^{kNN} : x \mapsto \mathbb{1}\{\hat{\eta}(x) \geq 1/2\}$  = plugin for Bayes classifier with estimator

$$\hat{\eta}(x) = \frac{1}{k} \sum_{j=1}^k Y_{(j)}(X)$$

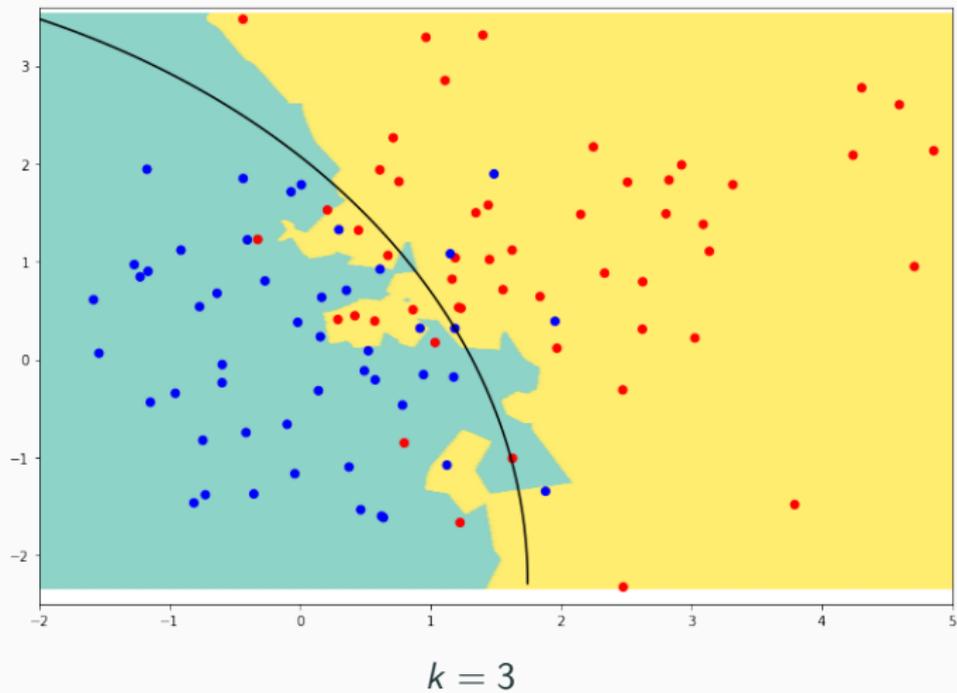
where

$$d(X_{(1)}(X), X) \leq d(X_{(2)}(X), X) \leq \dots \leq d(X_{(n)}(X), X) .$$

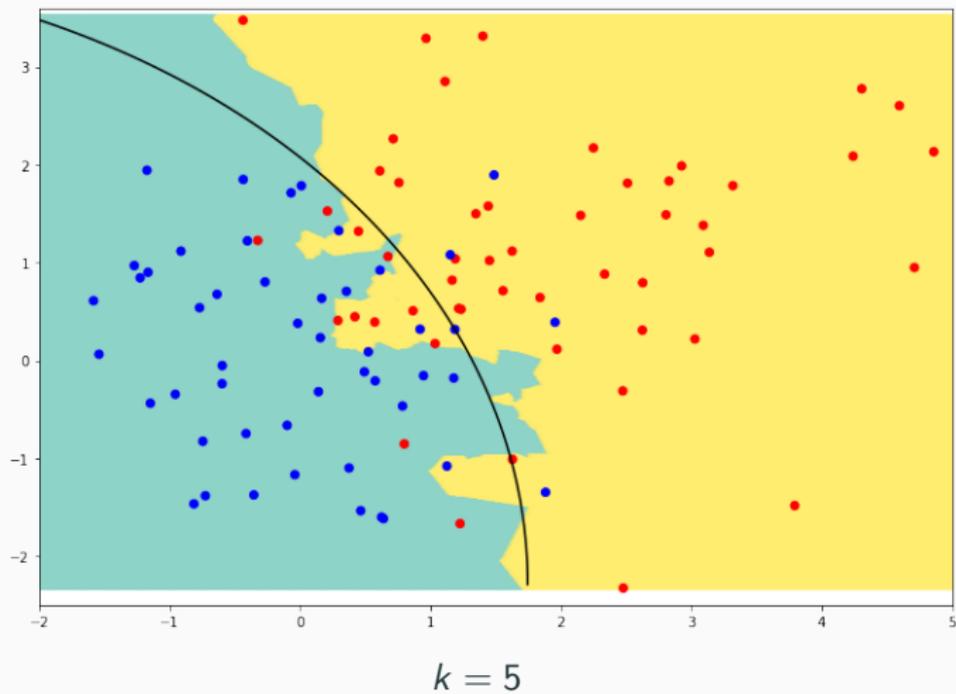
# More neighbors are better?



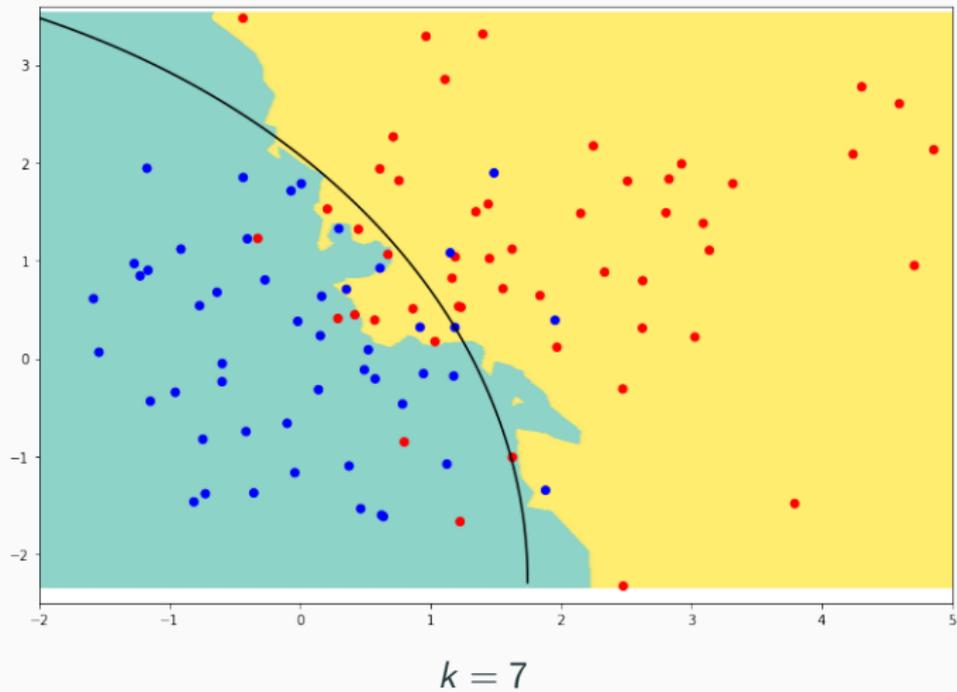
# More neighbors are better?



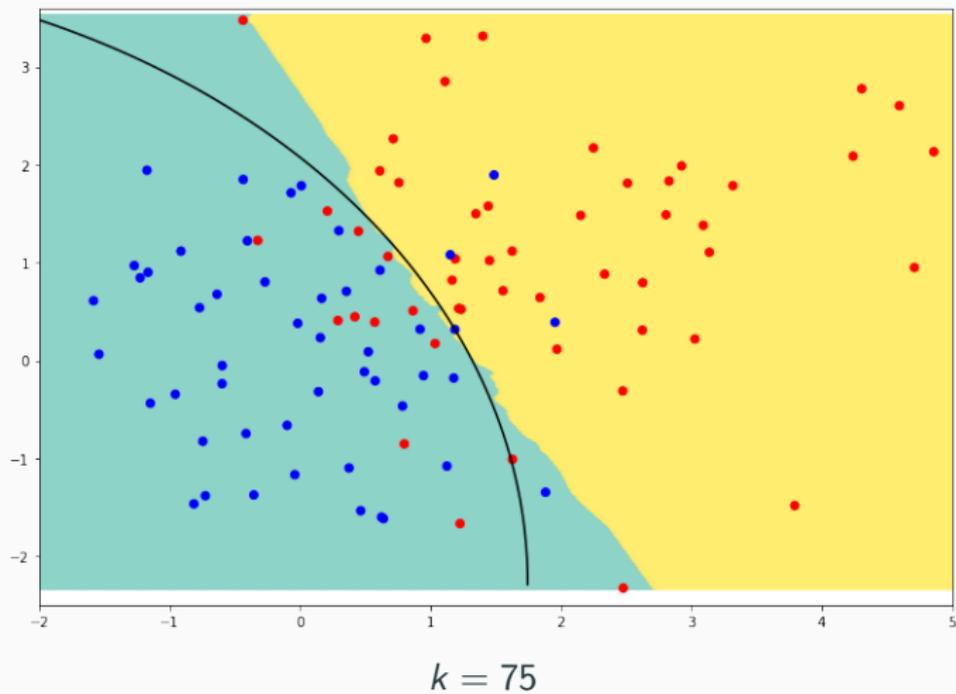
# More neighbors are better?



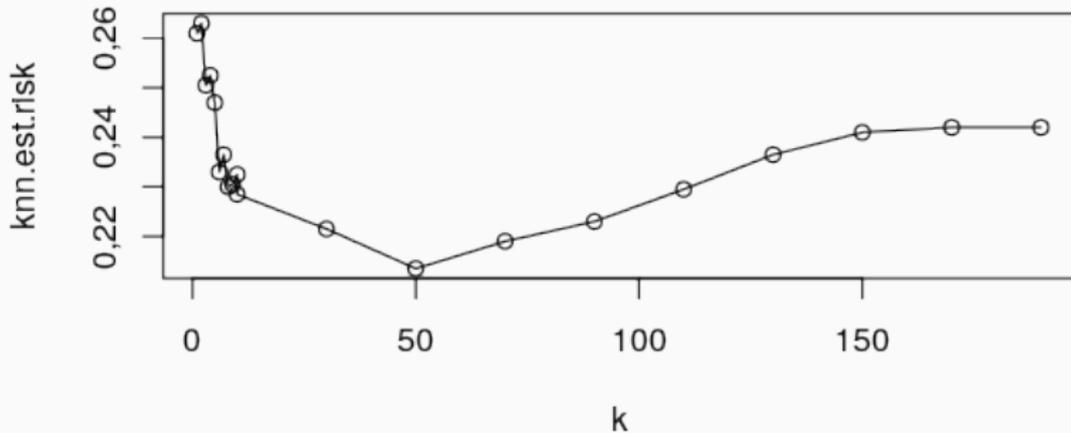
# More neighbors are better?



# More neighbors are better?



## Risque de k-NN en fonction du nombre de voisins



# Risk bound

Let  $\mathcal{C}_\epsilon$  be an  $\epsilon$ -covering of  $\mathcal{X}$ :

$$\forall x \in \mathcal{X}, \exists x' \in \mathcal{C}_\epsilon : d(x, x') \leq \epsilon .$$

## Excess risk for k-nearest-neighbours

If  $\eta$  is  $c$ -Lipschitz continuous:  $\forall x, x' \in \mathcal{X}, |\eta(x) - \eta(x')| \leq c d(x, x')$ ,  
then for all  $k \geq 2$  and all  $n \geq 1$ :

$$\begin{aligned} L(\hat{h}_n^{kNN}) - L(h^*) &\leq \frac{1}{\sqrt{ke}} + \frac{2k|\mathcal{C}_\epsilon|}{n} + 4c\epsilon \\ &\leq \frac{1}{\sqrt{ke}} + (2 + 4c) \left(\frac{\alpha k}{n}\right)^{\frac{1}{d+1}} \begin{cases} \text{for } \epsilon = \left(\frac{\alpha k}{n}\right)^{\frac{1}{d+1}}, \\ \text{if } |\mathcal{C}_\epsilon| \leq \alpha \epsilon^{-d} \end{cases} \\ &\leq (3 + 4c) \left(\frac{\alpha}{n}\right)^{\frac{1}{d+3}} \quad \text{for } k = \left(\frac{n}{\alpha}\right)^{\frac{2}{d+3}} . \end{aligned}$$

# Room for improvement

- Lower bound? in  $n^{-\frac{1}{d}}$ .
- Margin conditions
  - ⇒ fast rates
- More regularity?
  - ⇒ weighted nearest neighbors
- Is regularity required everywhere?
  - ⇒ What matters are the balls of mass  $\approx k/n$  near the decision boundary.
- 2 "parameters":
  - obvious: the number of neighbors  $k$  (bias-variance tradeoff)
  - hidden: the distance  $d$  (real problem)

# Curse of dimensionality: No free lunch theorem

## Theorem

Let  $c > 1$  be a Lipschitz constant. Let  $A$  be any learning algorithm for binary classification over a domain  $\mathcal{X} = [0, 1]^d$ . If the training set size is  $n \leq (c + 1)^d/2$ , then there exists a distribution  $\mathcal{D}$  over  $[0, 1]^d \times \{0, 1\}$  such that:

- $\eta(x)$  is  $c$ -Lipschitz;
- the Bayes error of the distribution is 0;
- with probability at least  $1/7$  over the choice of  $S_n \sim \mathcal{D}^{\otimes n}$ ,

$$L_{\mathcal{D}}(A(S_n)) \geq \frac{1}{8}.$$

# Empirical Risk Minimization

---

## Going empirical

Idea for every candidate rule  $h$  in an *hypothesis class*  $\mathcal{H}$ , replace the unknown risk

$$L_{\mathcal{D}}(h) = \mathbb{P}_{(X,Y) \sim \mathcal{D}}(h(X) \neq Y)$$

by the computable *empirical risk*

$$L_{S_n}(h) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{h(X_i) \neq Y_i\}$$

and use some *uniform law of large numbers*:

$$\mathbb{P}_{\mathcal{D}} \left( \sup_{h \in \mathcal{H}} |L_{S_n}(h) - L_{\mathcal{D}}(h)| > c \sqrt{\frac{D_{\mathcal{H}} \log(n) + \log \frac{1}{\delta}}{n}} \right) \leq \delta$$

where  $D_{\mathcal{H}}$  is the *Vapnik-Chervonenkis dimension* of  $\mathcal{H}$ .

# Empirical Risk minimization

*Uniform law of large numbers:*

$$\mathbb{P}_D \left( \sup_{h \in \mathcal{H}} |L_{S_n}(h) - L_D(h)| > c \sqrt{\frac{D_{\mathcal{H}} \log(n) + \log \frac{1}{\delta}}{n}} \right) \leq \delta .$$

→ *Empirical Risk Minimizer:*

$$\hat{h}_n = \arg \min_{h \in \mathcal{H}} L_{S_n}(h) .$$

Good if

- the class  $\mathcal{H}$  is not too large
- the number  $n$  of examples is large enough

so as to ensure that  $c \sqrt{\frac{D_{\mathcal{H}} \log(n) + \log \frac{1}{\delta}}{n}} \leq \epsilon$ .

→ *Sample complexity* = number of examples required to have an  $\epsilon$ -optimal rule in the hypothesis class  $\mathcal{H} = O\left(\frac{D_{\mathcal{H}}}{\epsilon^2}\right)$  .

# The class of halfspaces

## Definition

The class of linear (affine) functions on  $\mathcal{X} = \mathbb{R}^d$  is defined as

$$L_d = \{h_{w,b} : w \in \mathbb{R}^d, b \in \mathbb{R}\}, \quad \text{where } h_{w,b}(x) = \langle w, x \rangle + b.$$

The hypothesis class of halfspaces for binary classification is defined as

$$\mathcal{H}S_d = \text{sign} \circ L_d = \left\{x \mapsto \text{sign}(h_{w,b}(x)) : h_{w,b} \in L_d\right\}$$

where  $\text{sign}(u) = \mathbb{1}\{u \geq 0\} - \mathbb{1}\{u < 0\}$ . *Depth 1 neural networks.*

By taking  $\mathcal{X}' = \mathcal{X} \times \{1\}$  and  $d' = d + 1$ , we may omit the bias  $b$  and focus on functions  $h_w(x) = \langle w, x \rangle$ .

## Property

The VC-dimension of  $\mathcal{H}S_d$  is equal to  $d + 1$ .

Corollary: the class of halfspaces is learnable with *sample complexity*  
 $O\left(\frac{d+1+\log(1/\delta)}{\epsilon^2}\right)$ .

## Realizable case: Learning halfspaces with a linear program solver

*Realizable case:* there exists  $w^*$  such that  $\forall i \in \{1, \dots, n\}, y_i \langle w^*, x_i \rangle > 0$ .

Then there exists  $\bar{w} \in \mathbb{R}^d$  such that  $\forall i \in \{1, \dots, n\}, y_i \langle \bar{w}, x_i \rangle \geq 1$ : if we can find one, we have an ERM.

Let  $A \in \mathcal{M}_{n,d}(\mathbb{R})$  be defined by  $A_{i,j} = y_i x_{i,j}$ , and let  $v = (1, \dots, 1) \in \mathbb{R}^n$ . Then any solution of the linear program

$$\max_{w \in \mathbb{R}^d} \langle 0, w \rangle \quad \text{subject to} \quad Aw \geq v$$

is an ERM. It can thus be computed in polynomial time.

# Rosenblatt's Perceptron algorithm

---

**Algorithm:** Batch Perceptron

---

**Data:** training set  $(x_1, y_1), \dots, (x_n, y_n)$

- 1  $w_0 \leftarrow (0, \dots, 0)$
  - 2  $t \geq 0$
  - 3 **while**  $\exists i_t : y_{i_t} \langle w_t, x_{i_t} \rangle \leq 0$  **do**
  - 4      $w_{t+1} \leftarrow w_t + y_{i_t} \frac{x_{i_t}}{\|x_{i_t}\|}$
  - 5      $t \leftarrow t + 1$
  - 6 **return**  $w_t$
- 

Each updates helps reaching the solution, since

$$y_{i_t} \langle w_{t+1}, x_{i_t} \rangle = y_{i_t} \left\langle w_t + y_{i_t} \frac{x_{i_t}}{\|x_{i_t}\|}, x_{i_t} \right\rangle = y_{i_t} \langle w_t, x_{i_t} \rangle + \|x_{i_t}\|.$$

Relates to a coordinate descent (stepsize does not matter).

# Convergence of the Perceptron algorithm

## Theorem

Assume that the dataset  $S_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$  is linearly separable and let the *separation margin*  $\gamma$  be defined as:

$$\gamma = \max_{w \in \mathbb{R}^d: \|w\|=1} \min_{1 \leq i \leq n} \frac{y_i \langle w, x_i \rangle}{\|x_i\|}.$$

Then the perceptron algorithm stops after at most  $1/\gamma^2$  iterations.

**Proof:** Let  $w^*$  be such that  $\forall 1 \leq i \leq n, \frac{y_i \langle w^*, x_i \rangle}{\|x_i\|} \geq \gamma$ .

- If iteration  $t$  is necessary, then

$$\langle w^*, w_{t+1} - w_t \rangle = y_{i_t} \left\langle w^*, \frac{x_{i_t}}{\|x_{i_t}\|} \right\rangle \geq \gamma \quad \text{and hence } \langle w^*, w_t \rangle \geq \gamma t.$$

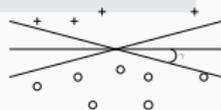
- If iteration  $t$  is necessary, then

$$\|w_{t+1}\|^2 = \left\| w_t + y_{i_t} \frac{x_{i_t}}{\|x_{i_t}\|} \right\|^2 = \|w_t\|^2 + \underbrace{\frac{2y_{i_t} \langle w_t, x_{i_t} \rangle}{\|x_{i_t}\|}}_{\leq 0} + y_{i_t}^2 \leq \|w_t\|^2 + 1$$

and hence  $\|w_t\|^2 \leq t$ , or  $\|w_t\| \leq \sqrt{t}$ .

- As a consequence, the algorithm iterates at least  $t$  times if

$$\gamma t \leq \langle w^*, w_t \rangle \leq \|w_t\| \leq \sqrt{t} \quad \implies \quad t \leq \frac{1}{\gamma^2}.$$



In the worst case, the number of iterations can be exponentially large in the dimension  $d$ . Usually, it converges quite fast. If  $\forall i, \|x_i\| = 1$ ,  $\gamma = d(S, D)$  where  $D = \{x : \langle w^*, x \rangle = 0\}$ .

# Computational difficulty of agnostic learning, and surrogates

## NP-hardness of computing the ERM for halfspaces

Computing an ERM in the agnostic case is NP-hard.

See *On the difficulty of approximately maximizing agreements*, by Ben-David, Eiron and Long.

Since the 0-1 loss

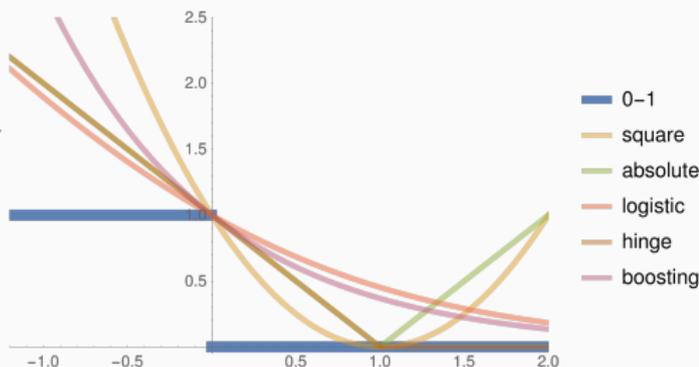
$$L_{S_n}(h_w) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{y_i \langle w, x_i \rangle < 0\}$$

is intractable to minimize in the agnostic case, one may consider *surrogate* loss functions

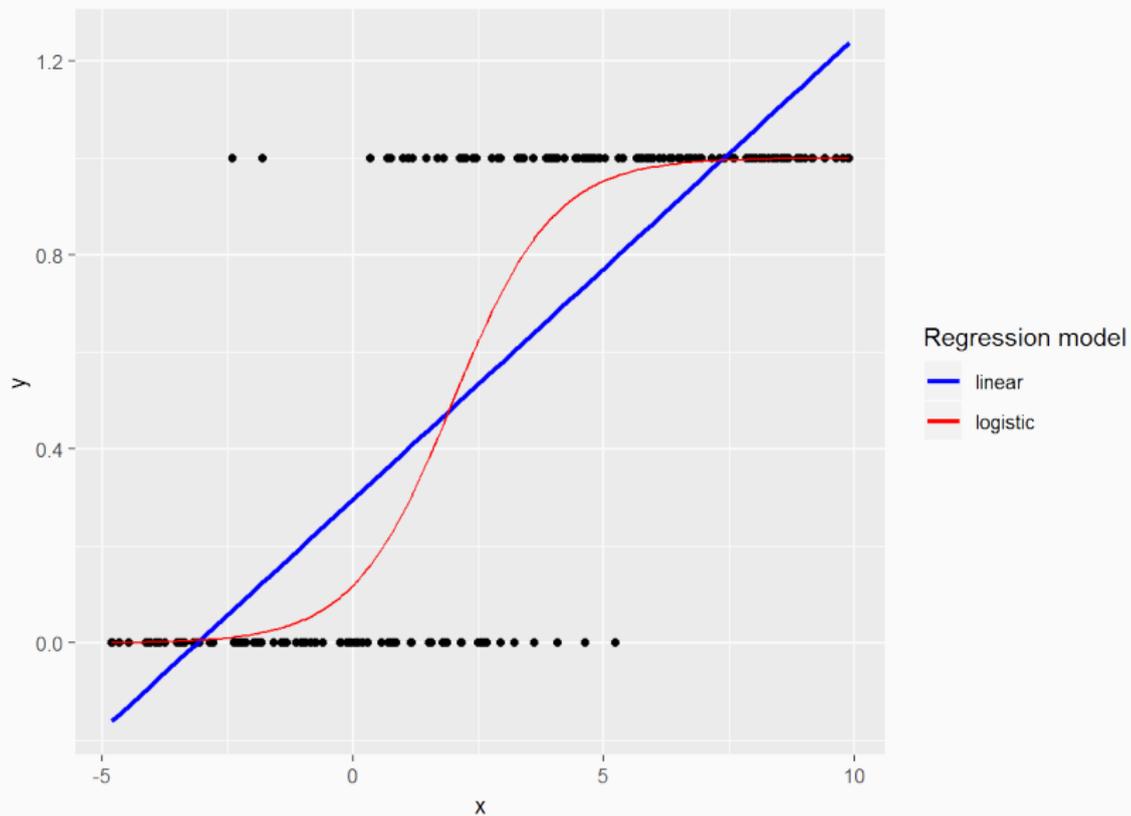
$$L_{S_n}(h_w) = \frac{1}{n} \sum_{i=1}^n \ell(y_i \langle w, x_i \rangle),$$

where the loss function  $\ell : \mathbb{R} \rightarrow \mathbb{R}^+$

- dominates the function  $\mathbb{1}\{u < 0\}$ ,
- and leads to a "simple" optimization problem (e.g. *convex*).



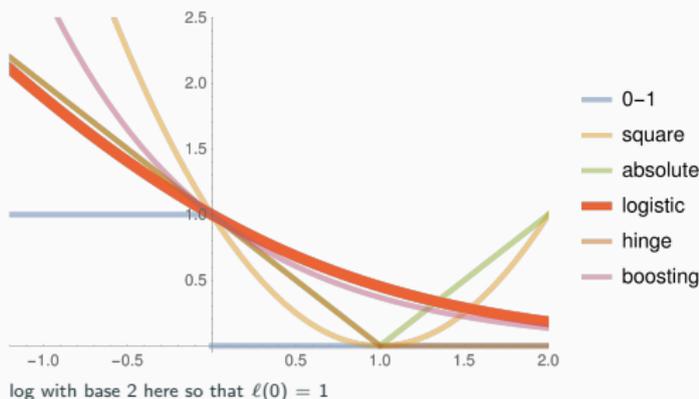
# Logistic Regression



# Logistic loss $\mathcal{Y} = \{-1, 1\}$

Statistics: "logistic regression":

$$P_w(Y = y|X = x) = \frac{1}{1 + \exp(-y \langle w, x \rangle)}$$



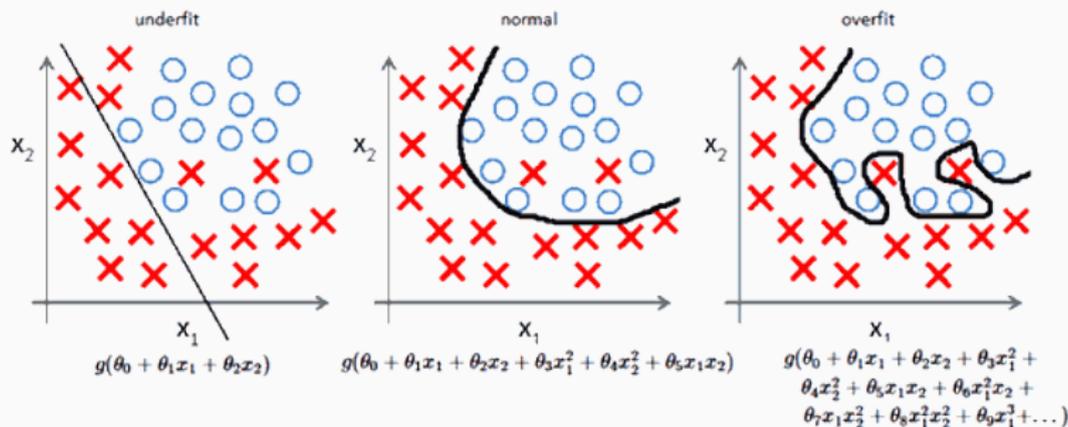
$$L_S(h_w) = \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-y_i \langle w, x_i \rangle)) ,$$

Convex minimization problem, can be solved by Newton's algorithm (in small dimension) or stochastic gradient descent (in higher dimension).

# Structural Risk minimization

What if  $\mathcal{H} = \bigcup_{d=1}^{\infty} \mathcal{H}_d$ , with  $\mathcal{H}_d \subset \mathcal{H}_{d+1}$ ?

→ empirical risk minimization fails



→ *structural* risk minimization:

$$\hat{h}_n = \arg \min_{d \geq 1, h \in \mathcal{H}_d} L_{S_n}(h) + D_{\mathcal{H}_d} \log(n).$$

# Support Vector Machines

---

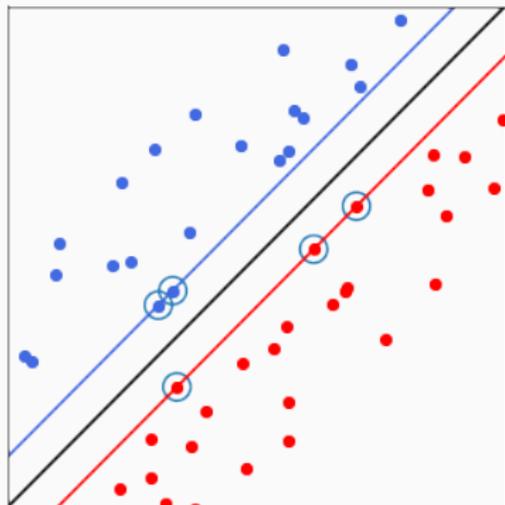
# Margin for linear separation

- Training sample  $S_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ , where  $x_i \in \mathbb{R}^d$  and  $y_i \in \{\pm 1\}$ .
- Linearly separable if there exists a halfspace  $h = (w, b)$  such that  $\forall i, y_i = \text{sign}(\langle w, x_i \rangle + b)$ .
- What is the best separating hyperplane for generalization?

## Distance to hyperplane

If  $\|w\| = 1$ , then the distance from  $x$  to the hyperplane  $h = (w, b)$  is  $d(x, \mathcal{H}) = |\langle w, x \rangle + b|$ .

**Proof:** Check that  $\min \{\|x - v\|^2 : v \in h\}$  is reached at  $v = x - (\langle w, x \rangle + b)w$ .



Formulation 1:

$$\arg \max_{(w,b): \|w\|=1} \min_{1 \leq i \leq m} |\langle w, x_i \rangle + b| \quad \text{such that } \forall i, y_i (\langle w, x_i \rangle + b) > 0.$$

Formulation 2:

$$\min_{w,b} \|w\|^2 \quad \text{such that } \forall i, y_i (\langle w, x_i \rangle + b) \geq 1.$$

Remark:  $b$  is not penalized.

## Proposition

The two formulations are equivalent.

**Proof:** if  $(w_0, b_0)$  is the solution of Formulation 2, then  $\hat{w} = \frac{w_0}{\|w_0\|}$ ,  $\hat{b} = \frac{b_0}{\|w_0\|}$  is a solution of Formulation 1: if  $(w^*, b^*)$  is another solution, then letting  $\gamma^* = \min_{1 \leq i \leq m} y_i (\langle w^*, x_i \rangle + b^*)$  we see that  $(\frac{w^*}{\gamma^*}, \frac{b^*}{\gamma^*})$  satisfies the constraint of Formulation 2, hence  $\|w_0\| \leq \frac{\|w^*\|}{\gamma^*} = \frac{1}{\gamma^*}$  and thus  $\min_{1 \leq i \leq m} |\langle \hat{w}, x_i \rangle + \hat{b}| = \frac{1}{\|w_0\|} \geq \gamma^*$ .

# Sample Complexity

## Definition

A distribution  $\mathcal{D}$  over  $\mathbb{R}^d \times \{\pm 1\}$  is *separable with a  $(\gamma, \rho)$ -margin* if there exists  $(w^*, b^*)$  such that  $\|w^*\| = 1$  and with probability 1 on a pair  $(X, Y) \sim \mathcal{D}$ , it holds that  $\|X\| \leq \rho$  and  $Y(\langle w^*, X \rangle + b) \geq \gamma$ .

Remark: by multiplying the  $x_i$  by  $\alpha$ , the margin is multiplied by  $\alpha$ .

## Theorem

For any distribution  $\mathcal{D}$  over  $\mathbb{R}^d \times \{\pm 1\}$  that satisfies the  $(\gamma, \rho)$ -separability with margin assumption using a homogenous halfspace, with probability at least  $1 - \delta$  over the training set of size  $n$  the 0 – 1 loss of the output of Hard-SVM is at most

$$\sqrt{\frac{4(\rho/\gamma)^2}{n}} + \sqrt{\frac{2 \log(2/\delta)}{n}}.$$

Remark: depends on dimension  $d$  only thru  $\rho$  and  $\gamma$ .

When the data is not linearly separable, allow *slack variables*  $\xi_i$ :

$$\min_{w,b,\xi} \lambda \|w\|^2 + \frac{1}{n} \sum_{i=1}^n \xi_i \quad \text{such that } \forall i, y_i (\langle w, x_i \rangle + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

$$= \min_{w,b} \lambda \|w\|^2 + L_{S_n}^{\text{hinge}}(w, b) \quad \text{where } \ell^{\text{hinge}}(u) = \max(0, 1 - u).$$

## Theorem

Let  $D$  be a distribution over  $B(0, \rho) \times \{\pm 1\}$ . If  $\mathcal{A}_n(S_n)$  is the output of the soft-SVM algorithm on the sample  $S$  of  $D$  of size  $n$ ,

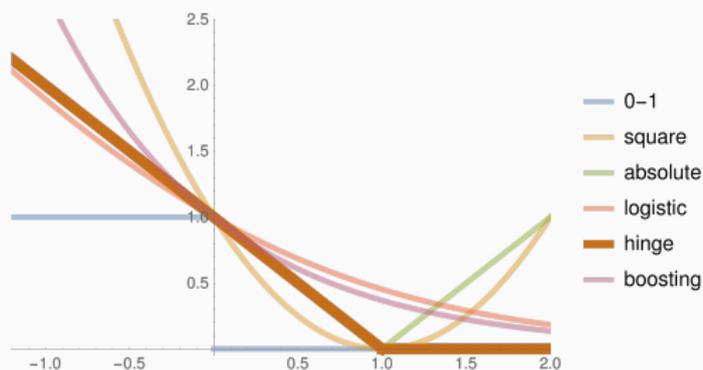
$$\mathbb{E} \left[ L_D^{0-1}(\mathcal{A}_n(S_n)) \right] \leq \mathbb{E} \left[ L_D^{\text{hinge}}(\mathcal{A}_n(S_n)) \right] \leq \inf_u L_D^{\text{hinge}}(u) + \lambda \|u\|^2 + \frac{2\rho^2}{\lambda n}.$$

For every  $B > 0$ , setting  $\lambda = \sqrt{\frac{2\rho^2}{B^2 n}}$  yields:

$$\mathbb{E} \left[ L_D^{0-1}(\mathcal{A}_n(S_n)) \right] \leq \mathbb{E} \left[ L_D^{\text{hinge}}(\mathcal{A}_n(S_n)) \right] \leq \inf_{w: \|w\| \leq B} L_D^{\text{hinge}}(w) + \sqrt{\frac{8\rho^2 B^2}{n}}.$$

# SVM as a Penalized Empirical Risk Minimizer

Margin maximization leads to



$$L_{S_n}^{\text{hinge}}(h_w) = \frac{1}{n} \sum_{i=1}^n \max \{0, 1 - y_i \langle w, x_i \rangle\},$$

convex but non-smooth minimization problem, used with a penalization term  $\lambda \|w\|^2$ .

## Dual Form of the SVM Optimization Problem

To simplify, we consider only the homogeneous case of hard-SVM. Let

$$g(w) = \max_{\alpha \in [0, +\infty)^n} \sum_{i=1}^n \alpha_i (1 - y_i \langle w, x_i \rangle) = \begin{cases} 0 & \text{if } \forall i, y_i \langle w, x_i \rangle \geq 1, \\ +\infty & \text{otherwise.} \end{cases}$$

Then the hard-SVM problem is equivalent to

$$\begin{aligned} \min_{w: \forall i, y_i \langle w, x_i \rangle \geq 1} \frac{1}{2} \|w\|^2 &= \min_w \frac{1}{2} \|w\|^2 + g(w) \\ &= \min_w \max_{\alpha \in [0, +\infty)^n} \frac{1}{2} \|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i \langle w, x_i \rangle) \\ &\stackrel{\text{min-max thm}}{=} \max_{\alpha \in [0, +\infty)^n} \min_w \frac{1}{2} \|w\|^2 + \sum_{i=1}^n \alpha_i (1 - y_i \langle w, x_i \rangle). \end{aligned}$$

The inner min is reached at  $w = \sum_{i=1}^n \alpha_i y_i x_i$  and can thus be written as

$$\max_{\alpha \in \mathbb{R}^n, \alpha \geq 0} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{1 \leq i, j \leq n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle.$$

Still for the homogeneous case of hard-SVM:

## Property

Let  $w_0$  be a solution of and let  $I = \{i : |\langle w_0, x_i \rangle| = 1\}$ . There exist  $\alpha_1, \dots, \alpha_n$  such that

$$w_0 = \sum_{i \in I} \alpha_i x_i .$$

The dual problem involves the  $x_i$  only thru scalar products  $\langle x_i, x_j \rangle$ .

It is of size  $n$  (independent of the dimension  $d$ ).

These computations can be extended to the non-homogeneous soft-SVM

→ **Kernel trick.**

# Numerically solving Soft-SVM

$f(w) = \frac{\lambda}{2} \|w\|^2 + L_S^{\text{hinge}}(w)$  is  $\lambda$ -strongly convex.

→ Stochastic Gradient Descent with learning rate  $1/(\lambda t)$ . Stochastic subgradient of  $L_S^{\text{hinge}}(w)$ :  $v_t = -y_{l_t} x_{l_t} \mathbb{1}\{y_{l_t} \langle w, x_{l_t} \rangle < 1\}$ .

$$w_{t+1} = w_t - \frac{1}{\lambda t} (\lambda w_t + v_t) = \frac{t-1}{t} w_t - \frac{1}{\lambda t} v_t = -\frac{1}{\lambda t} \sum_{i=1}^t v_i.$$

---

## Algorithm: SGD for Soft-SVM

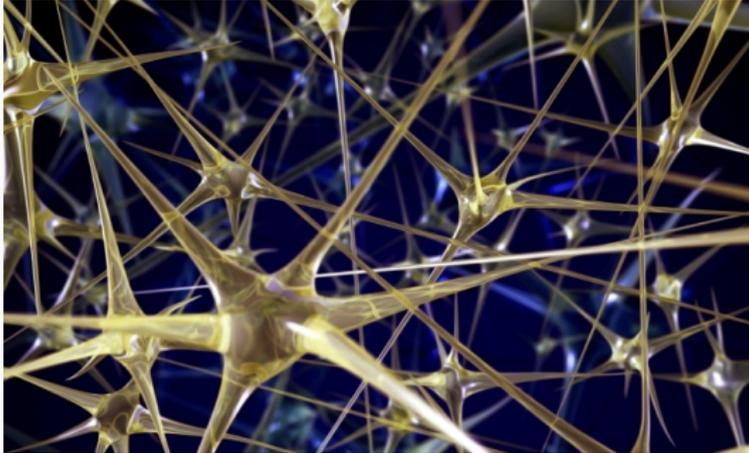
---

- 1 Set  $\theta_0 = 0$
- 2 **for**  $t = 0 \dots T - 1$  **do**
- 3     Let  $w_t = \frac{1}{\lambda t} \theta_t$
- 4     Pick  $l_t \sim \mathcal{U}(\{1, \dots, n\})$
- 5     **if**  $y_{l_t} \langle w_t, x_{l_t} \rangle < 1$  **then**
- 6          $\theta_{t+1} \leftarrow \theta_t + y_{l_t} x_{l_t}$
- 7     **else**
- 8          $\theta_{t+1} \leftarrow \theta_t$
- 9 **return**  $\bar{w}_T = \frac{1}{T} \sum_{t=0}^{T-1} w_t$

# Neural Networks

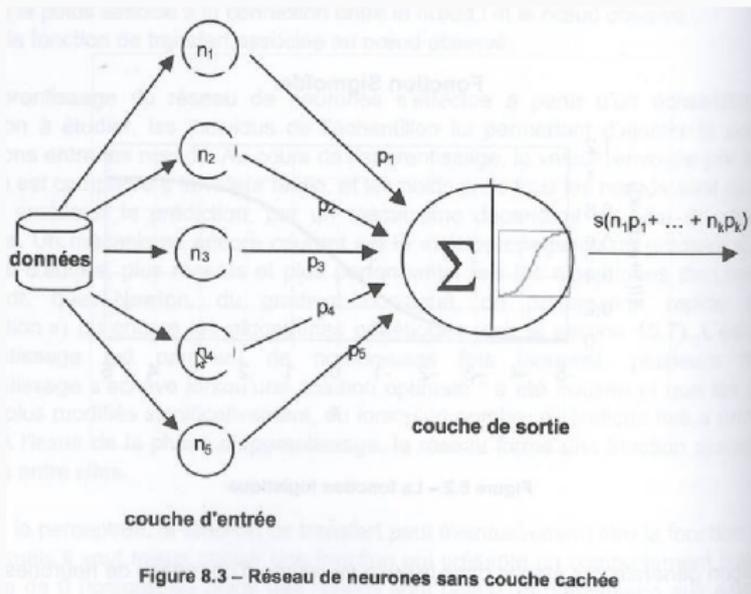
---

# One-layer network



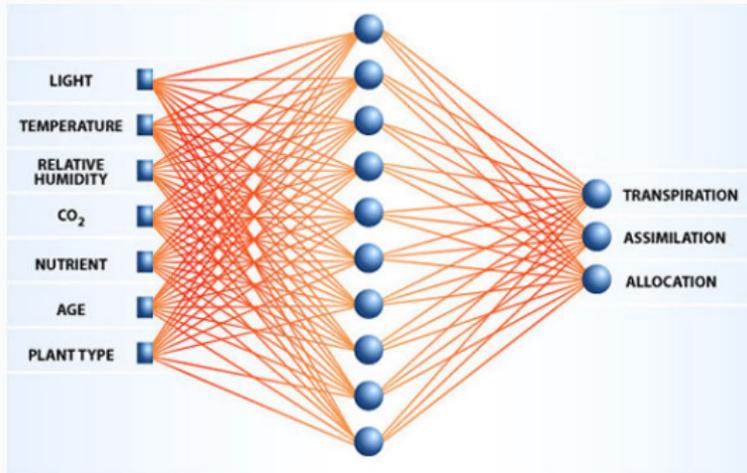
Src: <http://insanedevelop.co.uk/open-cranium/>

# One-layer network



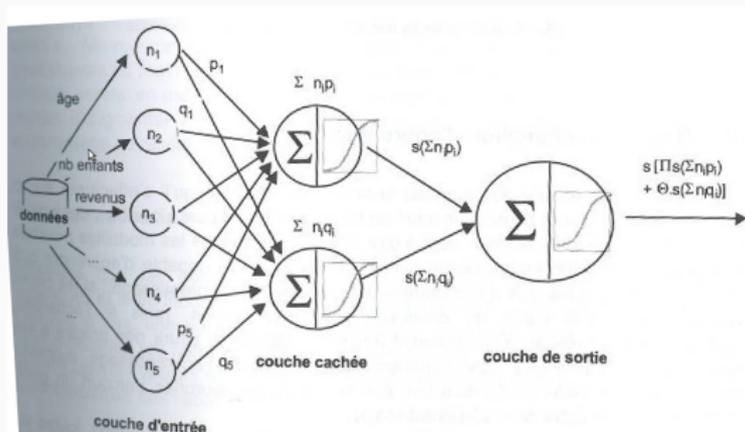
Src: [Tufféry, Data Mining et Informatique Dcisionnelle]

# One-layer network



Src: <http://www.makhfi.com>

# Two-layer network



Src: [Tufféry, Data Mining et Informatique Dcisionnelle]

# Profound ideas and tricks

- Convolutional networks
- Max-pooling
- Dropout
- Data augmentation
- GANs
- Representation learning
- Self-learning (ex: classify against rotations)

# The three main theoretical challenges of deep learning

- **Expressive power of DNN:** why are the function we are interested in so well approximated by (deep convolutive) neural networks?
- **Success of naive optimisation:** why does gradient descent lead to a good local minimum?
- **Generalization miracle** why is there no overfitting with so many parameters?

