HOMEWORK 3

$YangGao$
9083410275

Solution 1

Solution 1.1

(a) Regression, since CEO salary is a continuous variable.
n = 500
p = 3 (profit, number of employees, industry)
(b) Classification, as there are 2 possible outcomes, or 2 classes.
n = 20
p = 13 (price, marketing budget, competition price + 10 other)
(c) Regression, since percent change is a continuous variable.
n = 52 (weeks in year 2012)
p = 3 (% change in US market, % change in British market, % change in German market)

Solution 1.2

(a) distances: [3, 2, 3.1622776601683795, 2.23606797749979, 1.4142135623730951, 1.7320508075688772]
(b) Prediction for K=1 is Green, as the first closest point to [0,0,0] is [-1,0,1], and that point's target is Green.
(c) Prediction for K=3 is Red, as the 3 closest points to [0,0,0] have the labels Green, Red, Red. Red is most popular, so the results is Red.

Solution 1.3

(a) 1/10
(b) 0.1*0.1 = 1/100
(c) $\frac{1}{10}^{100} = 1e-100$
(d) Since number of close neighbours decrease exponentially with the increase of p, when p is large there will be proportionately very few neighbours near the test point. This may be ok if we have many training samples that are well distributed, otherwise KNN will have few good close neighbours to base a good prediction on.
(e) Let x be the length of each side of the hypercube: $x^p = 0.1$
$x = \sqrt[p]{0.1}$

p	length
1	0.1
2	0.316
100	0.977

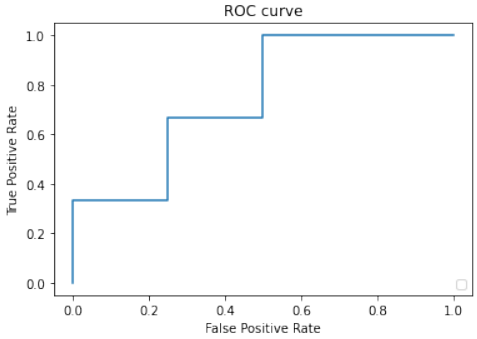
As p increases, we have to cover more and more of the sample space to find closest neighbours. This implies that KNN is not a suitable method as number of features becomes large.

Solution 1.4

(a) Accuracy = (8+974)/(8+2+16+974) = 98.2%
(b) Precision = TP/(TP+FP) = 8/(8+16) = 33.3%
(c) Recall = TP/(TP + FN) = 8/(8+2) = 80%

Solution 1.5

(a)



Q5a

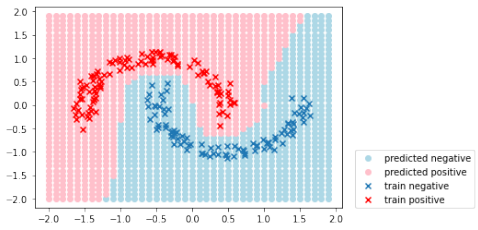
(b) We want the True positive rate to be 1 and False positive rate to be 0 for optimal performance. Based on the ROC curve, we see that no threshold perfectly satisfies this condition, but we can choose a threshold that optimizes TP and FP. At threshold = 0.8, we have approximately TP=0.6 and FP = 0.3, which is a good threshold. I chose to not let FP be very high, as I hate it when important emails go into spam, but I also don't want to receive spam.

Solution 1.6

(a)
[-0.50, -1.50, -1.00]
(b) $\theta^1 = [0.05, 0.15, 0.10]$

Solution 2

Solution 2.1



Q2.1

Solution 2.2

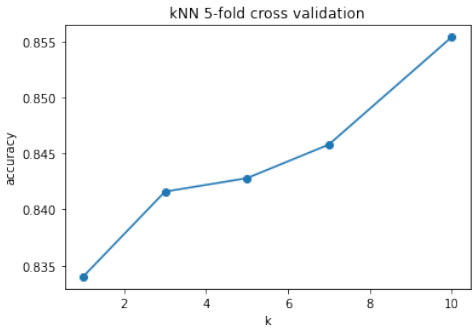
Fold	accuracy	precision	recall
0	0.824	0.6526610644257703	0.8175438596491228
1	0.855	0.6896551724137931	0.8664259927797834
2	0.861	0.7217125382262997	0.8309859154929577
3	0.854	0.7215568862275449	0.8197278911564626
4	0.776	0.6067708333333334	0.761437908496732

Solution 2.3

Fold	accuracy	precision	recall
0	0.971	0.9266666666666666	0.9754385964912281
1	0.973	0.9401408450704225	0.9638989169675091
2	0.969	0.9347079037800687	0.9577464788732394
3	0.958	0.9285714285714286	0.9285714285714286
4	0.942	0.8924050632911392	0.9215686274509803

Solution 2.4

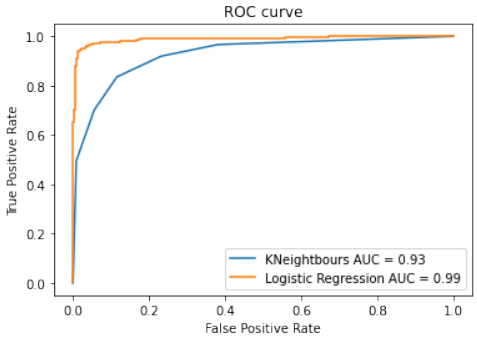
Average accuracy vs. k:



Q2.4

NN	accuracy
1	0.8340
3	0.8416
5	0.8428
7	0.8458
10	0.8554

Solution 2.5



Q2.5