CSCE633 Homework 02

Lu Sun

September 2020

1 (0.5 points) Which of the following problems are best suited for machine learning?

Answer: BD

A: Don't have clear formula (way to measure) a number is prime or not.

B: Have task, detecting fraud, measurement method, credit card charges

C: The time can be computed according to fixed known formula. No need to generate new formula.

D: A prediction problem in machine learning.

2 (1 points) Please select the correct answer(s).

Answer: C

K = 3.

For x_1, $||x_1 - x_2||_2 = 1, ||x_1 - x_3||_2 = \sqrt{2}, ||x_1 - x_4||_2 = 1, ||x_1 - x_5||_2 = \frac{\sqrt{2}}{2}$,nearest K point: x_2, x_4, x_5, then $y_1^{pred} = 1 = y_1$, true, $accuracy_1 = 1$

For x_2, $||x_1 - x_2||_2 = 1, ||x_2 - x_3||_2 = 1, ||x_2 - x_4||_2 = \sqrt{2}, ||x_2 - x_5||_2 = \frac{\sqrt{2}}{2}$,nearest K point: x_1, x_3, x_5, then $y_2^{pred} = 1 = y_2$, true, $accuracy_2 = 1$

For x_3, $||x_1 - x_3||_2 = \sqrt{2}, ||x_2 - x_3||_2 = 1, ||x_3 - x_4||_2 = 1, ||x_3 - x_5||_2 = \frac{\sqrt{2}}{2}$,nearest K point: x_2, x_4, x_5, then $y_3^{pred} = 1 = y_3$, true, $accuracy_3 = 1$

For x_4, $||x_1 - x_4||_2 = 1, ||x_4 - x_2||_2 = \sqrt{2}, ||x_4 - x_3||_2 = 1, ||x_4 - x_5||_2 = \frac{\sqrt{2}}{2}$,nearest K point: x_1, x_3, x_5, then $y_4^{pred} = 1 = y_4$, true, $accuracy_4 = 1$

For x_5, $||x_1 - x_5||_2 = \frac{\sqrt{2}}{2}, ||x_5 - x_2||_2 = \frac{\sqrt{2}}{2}, ||x_5 - x_3||_2 = \frac{\sqrt{2}}{2}, ||x_5 - x_4||_2 = \frac{\sqrt{2}}{2}$,nearest K point: x_1, x_2, x_3, x_4, then $y_5^{pred} = 0 \neq y_5 = 0$, false, $accuracy_5 = 0$

Average accuracy $= \frac{4*1+0}{5} = 0.8$

3 (1 points) Please select the correct answer(s).

Answer: AD

$SSR := \sum_{n=1}^{N} (w_0 + w_1 x_n - y_n)^2$

$\frac{\partial SSR}{\partial w_0} = 2\sum_{n=1}^{N}(w_0+w_1x_n-y_n) = 2(w_0\sum_{n=1}^{N}1+w_1\sum_{n=1}^{N}x_n-\sum_{n=1}^{N}y_n) \overset{\{x_n,y_n\}_{n=1}^{N} have\ 0\ mean}{=} 2Nw_0 = 2*10*w_0$

$\frac{\partial SSR}{\partial w_1} = 2\sum_{n=1}^{N}(w_0+w_1x_n-y_n)x_n = 2(w_0\sum_{n=1}^{N}x_n+w_1\sum_{n=1}^{N}x_n^2-\sum_{n=1}^{N}x_ny_n) \overset{\sum_{n=1}^{N}x_n=0,\ \sum_{n=1}^{N}x_n^2=20,\ \sum_{n=1}^{N}x_ny_n=1}{=} 2*(20w_1-1)$

So the true answer should be (A) (C) (D)'s w_0 and (A) and (D)'s w_1. For (A), set $2*\alpha(k)=\beta(k)$. For (D), set $\alpha(k)=10*\beta(k)$.

4 (1 point) Which of the following loss functions would not work well with gradient descent?

Answer: B

A: concave function, gradient descent method can work well to find global maximum.

C: convex function, gradient descent method can work well to find global minimum.

B: not convex or concave, exists saddle point, that gradient descent method doesn't work.

5 (2 points) The code should output the average accuracy of testing data over all three folds.

Answer:

```
#Input: X, y
#Output: ave_acc

def LP(X_train,y_train,X_test,y_test):
    W initialized as  D dim- zero vector
    max_iter = 10000
    t = 0
    while t< max_iter:
        if there exists x in X_train, y in y_train, such that ...
            sign(W^T x) !=y: #misclassified
            W = W + y * (W^T x)
        else:
            break
        t = t+1
    if t == max_iter:
        print("not linear seperable data set")
        return False
    predict = sign(W^T X_test)
    acc = number of (predict-y_test equal to 0 )  /  number of ...
        y_test
    return acc

```

```
Acc = [] #record for each all three folds' accuracy
for s in [1,2,3]:
    acc = LP(X\Xs,y\ys,Xs,ys)
    Acc.append(acc)
ave_acc = mean value of Acc
return ave_acc
```

6 (0.5 points) True / False : A classifier trained on less training data is less likely to overfit.

Answer: False

Less training data results in more likely overfitting.

7 (1 point) Would you accept or reject each paper?

Answer:

Reject (A): 1.may result in overfitting. 2.when need to predict, lower training error isn't useful.

Reject (B): regularization parameter is too specific and depends too much on the data set.

Reject (C): regularization parameter's setting should only related to training and validation data with cross-validation method or other method, not chosen with test data set. Otherwise, may result in overfitting.

Accept (D): hyper parameter regularization parameter gets from validation set and can result in better prediction on testing data set.

8 (1 point) What would be good choices of x_1 and x_2?

Answer: C

x in range of $[1, 900]$, $\sqrt{x}$ in range of $[1, 30]$, in order to make the two ranges x_1, x_2 to be the same (than the gradient's range or w_1, w_2 is similar), $x_1 = \frac{x}{30}, x_2 = \sqrt{x} or x_1 = x, x_2 = 30\sqrt(x)$. Namely, C or D.

In order to make w_0's range $(w_0 + w_1x_1 + w_2x_2 - y) * 1$ as closed as w_1 and w_2, x_1 and x_2 should choose the range closest to range $[0, 1]$ corresponding to 1, namely, C is better.

9 (1 point) Which of the following classifier(s) would be able to successfully classify the following data?

Answer: A

A: can generate non-linear boundary.

C: need the data set to be linear separable.

B: can only result in accuracy to be 0, in each direction, there exist both red and blue classes.

10 (1 point) Which of the following model(s) would you suggest to the recruiter to use and why?

Answer: B

A: more like linear regression and ignore high performance prior experience people, but hgih performance and high prior experience people is the exact what recruiter wants.

B: fits well for data and has reasonable predict.

C: overfitting