Practice questions MAIR 2019 (machine learning)

24 September 2019

1 True or False?

- **[1] K-nearest neighbors is more efficient at test time than logistic regression**
  False, with KNN you need to compute the distance between the test instance and each training instance (assuming a naive implementation).

- **[2] 3-Nearest neighbors is more robust to outliers than 1-Nearest neighbors**
  True, see the example in the slides.

- **[3] A decision tree can always achieve 100% accuracy on the training set if there are no constraints when growing the tree (e.g. its depth is not limited) and there is no noise in the training set.**
  True. An extreme case would be a leaf for each training example.

2 Development of ML models

[4] Why can't you tune K of K-nearest-neighbors by looking at the training error?

The best K based on the training error would always be K=1. K has to be set on a development set (e.g. using cross validation).

[5] Your logistic regression model has a very low error rate on the training set but a high error rate on the development set. What is going on? How could you improve your model?

This is a case of overfitting. Add a regularization term (L1 or L2). Or, if you are already using regularization, increase λ.

	True label: Spam	**True label: Not spam**
Predicted label: Spam	10	15
Predicted label: Not spam	20	90

Table 1: Spam classification

[6] Compute the precision, recall and accuracy for the 'spam' category based on the confusion matrix shown in Table 1

Precision: 10/25
Recall: 10/30
Accuracy: 100/135

[7] You are asked to develop a model to predict the number of books that will be sold in the first week after a book is released.

[7a] This is a [regression/classification] problem.

This is a regression problem.

[7b] List 5 features that you could use.

For example, the author of the book, the genre of the book, the advertising budget, the number of book stores selling the book, the release date.

3 Distances

[8] What is the cosine similarity between a=[1, 1, 4] and b=[0, 1, 1]?

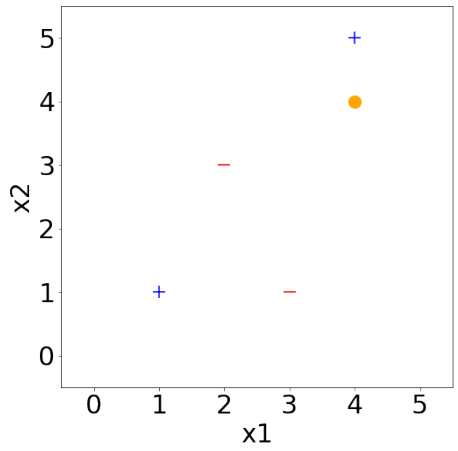
$5/(\sqrt{18} * \sqrt{2}) = 0.833$

[9] Given two vectors a=[1, 2, 3] and b=[1, 3, 5], compute the L1 and the L2 norm

L1: $|1-1|+|2-3|+|3-5|=3$
L2: $\sqrt{(1-1)^2+(2-3)^2+(3-5)^2}=\sqrt{5}$

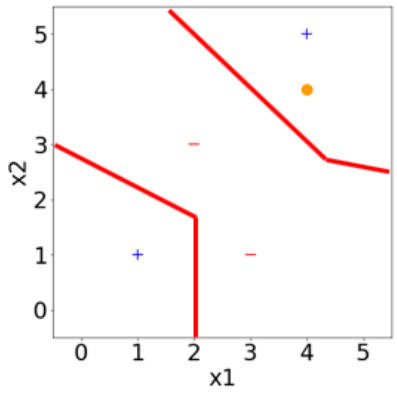
4 Nearest neighbors

[10a] The figure shows a dataset with two categories (+ and -). What would the classification be for the new point (the orange dot) using 1-NN? And using 3-NN?



Using 1-NN: + ; Using 3-NN: -

[10b] Sketch the decision boundary for 1NN (assume Euclidian distance)



5 Decision Trees

You have the following training data with 3 features (Humidity, Outlook, Wind) and a binary output (Y=0 or Y=1).

Y	Humidity	Outlook	Wind
0	High	Sunny	Strong
0	Normal	Rain	Weak
1	High	Overcast	Strong
1	High	Sunny	Weak
1	Normal	Rain	Weak

[11a] What is the entropy of Y? (i.e.. H(Y))

-log2(0.4) * 0.4 -log2(0.6) * 0.6 = 0.97

[11b] You fit a decision tree on the training data with no constraints (e.g. no maximum depth). How many instances would be misclassified? There is noise in the data: there are two instances with exactly the same features but different labels (Normal, Rain, Weak). These two instances would therefore end up in the same leaf. One instance would thus be misclassified.

6 Neural Networks

[12] ReLU What would the activation value be for a ReLU unit when you have the following input values: $x_1 = 1, x_2 = -1, x_3 = 2$ and the following weights $w_1 = 1, w_2 = 4, w_3 = 1$.

1*1 + -1 * 4 + 2 * 1 = -1.
max(-1,0) = 0

[13] Your neural network has a high error on the training set and similarly also a high error on the development set. One way to try to improve your neural network would be to increase the dropout probability. True or false?
False, your neural network is (assuming no bugs) underfitting. Instead it would be better to decrease the dropout probability.

7 Logistic Regression

[14] A logistic regression classifier returns $P(Y = 1|x) = 0.75$ for a binary classification problem. The correct label is Y=0. What is the cross entropy loss?

-log(1-0.75) = -log(0.25) = 1.386

[15] Let's say for a classification task we change the data by adding 3 features that are exact copies of one of the original features. Which classifier would be more impacted by this change: nearest-neighbors or logistic regression?

Nearest-neighbors as each feature is weighted equally. Instead, logistic regression can adjust the weights of individual features.