Information Mining - winter semester 2017

Exercise sheet 5

Excercise 1: Logistic Regression

Define logistic regression. Suppose that you have trained a logistic regression classifier, and it outputs on a new example x a prediction $h_w(x) = 0.4$. Which of the following statements are true:?

- $P(y = 1|x; w) = 0.6$
- $P(y = 0|x; w) = 0.4$
- $P(y = 1|x; w) = 0.4$
- $P(y = 0|x; w) = 0.6$

第一个。我猜的。不会。

Justify your answer.

Excercise 2: Undertanding

Explain in your words the issues high variance and high bias.

是数据上的还是分类器做的预测？

Excercise 3: k-nn

In the lecture we have seen that k-nn can be used for instance based classification. An instance is classified by its k neighbors (k is a positive integer, typically small). You discover high variance in your model and you suspect that your data has noise. How do you select your k to reduce the high variance problem?

select a higher k.

Excercise 4: K-means clustering

Explain the core idea of the K-means algorithm. How important is the selection of the seeds and what are the possible causes of the seed selection? How do you determine the k value (number of clusters)?

idea:
seed selection is important , a good select can shorten learning time use.
possible cause : get trapped in local minimum
determine k : combine performance and MDL principle.

Excercise 5: Understanding: Evaluation methods

(a) While classifying the following four outcomes can occur. Please shortly explain the meaning of these terms and give an example.

- true positive, TP
- true negative, TN
- false negative, FN
- false positive, FP

predict positive and it is right.
predict negative and it is right.
predict negative and it is wrong.
predict positive and it is wrong.

(b) How are the following measures defined? What do they define?

- accuracy
- mean-squared error
- precision and recall
- F1-measure

(TP+TN)/(TP+FP+TN+FN)
...
precision : TP/(TP+FP)
recall : TP/(TP+FN)
harmonic mean of precision and recall ,

(c) Which one of the above mentioned measures would be more applicable for the house price prediction? Shortly justify your answer.

MSE , because it is numeric , we cant use other 3 measures.