INF264 Introduction to machine learning

Model solutions and grading criteria

4.12.2019

1 Basic concepts

1. A model is overfitting if it performs well on training data but it has bad generalization performance (it performs poorly on unseen validation data). Bad generalization performance is a problem because the main goal in machine learning is generalization.

2. In dimensionality reduction, the goal is to represent data with less features while retaining as much information as possible. It can be useful, e.g., because it reduces the risk of overfitting, makes computations faster and enables easier visualization.

3. Typically, we are interested in detecting the minority class so measures like precision, recall, or F1 score are good choices. (If data is imbalanced then a classifier that always predicts the majority class gives good accuracy. Thus, accuracy is a problematic performance measure if the data is imbalanced.)

4. Information gain is used in decision tree learning to decide which variable to split. Informally, information gain is the difference between uncertainty before the split and uncertainty after the split. Formally, information gain of splitting variable x is $IG(x) = H(y) - H(y|x)$, where $H(y)$ is the entropy of the class label and $H(y|x)$ is the conditional entropy of y given x.

5. One can use non-linear basis functions to transform data to a high-dimensional space and perform linear regression in the new space.

Grading: max 2 points for each task. Full two points for mentioning all key points, 1 point for some correct elements. Deductions of 0.5-1 points for incorrect statements.

2 k-means clustering

Initial cluster centers: $\mu_1 = 2$ and $\mu_2 = 5$

Assign each data point to the closest cluster center. We get a partition $D_1 = \{0, 2, 3\}$ and $D_2\{5, 8\}$

Update cluster centers: $\mu_1 = \frac{0+2+3}{3} = 1\frac{2}{3}$ and $\mu_2 = \frac{5+8}{2} = 6\frac{1}{2}$

Assign data point to the closest cluster center. We get a partition $D_1 = \{0, 2, 3\}$ and $D_2\{5, 8\}$. We observe the the partition did not change. Thus, we can stop.

Grading:

The idea of k-means clustering 4p, Correct cluster assignment step 3p, Correct cluster mean update step 3p.

Deductions:

Numerical errors 0.5-1p.

Note that the task was to simulate Lloyd's algorithm. If you get the correct clustering using some other method then you will be penalized 4-7 points depending on how closely your clustering algorithm resembles Lloyd's algorithm.

3 Support vector machines

Decision boundary, margins and support vectors ar shown in Figure 1. The data are linearly separable so training accuracy is 1.

There are 6 data points so 6-fold cross validation always leaves one point to the validation set and the remaining 5 to the training set. Consider the classifier that is shown in Figure 1. We note that removing a non-support vector does not change the decision boundary. Thus, validation sets $(6, 3)$ and $(6, 1)$ give validation accuracy 1. Furthermore, data points $(1, 1)$, $(1, 2)$, and $(1, 3)$ lie on the same margin. Thus, removing one of them does not affect the decision boundary and validation sets $(1, 1)$, $(1, 2)$, and $(1, 3)$ give validation accuracy one. However, removing $(3, 2)$ changes the decision boundary. When $(3, 2)$ is in the validation data, the decision boundary is $x_1 = 3.5$ (all

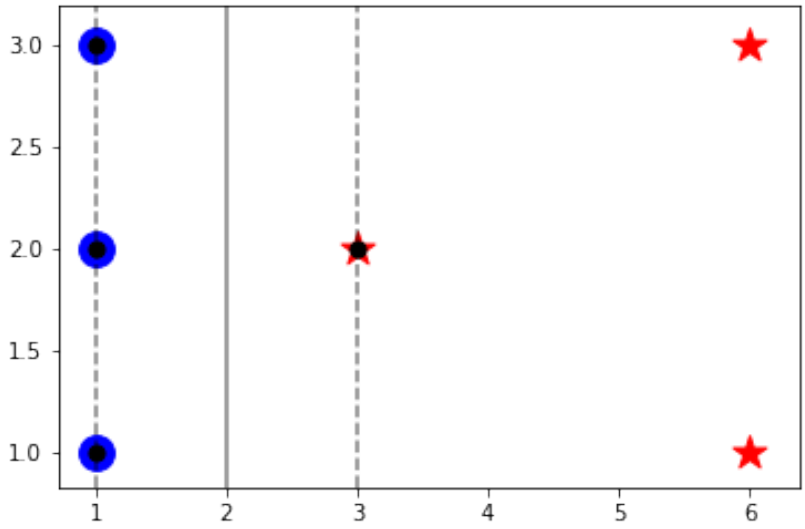

Figure 1: Decision boundary (solid line), margins (dashed lines) and support vectors (black dot in the middle of a marker).

points with $x_1 \leq 3.5$ are classified as -1) and $(3, 2)$ is misclassified giving validation accuracy 0.

Averaging over 6 validation sets, we get validation accuracy $\frac{5}{6}$

Grading:

3.1 (5p): Decision boundary correct 2p, margins correct 2p, support vectors correct 1p

3.2 (1p)

3.3 (4p): Idea of cross-validation correct 1p, SVMs trained correctly 2p, final validation accuracy computed correctly 1p

Points are deducted for missing details and incorrect claims

4 Bagging and random forests

Bagging is an ensemble method where one learns several classifiers using bootstrap samples and predicts by letting the classifiers vote. A bootstrap sample of a data set with n points is a n-point sample with replacement.

In bagging, the idea is to take several models that have low bias and high variance (that is, they overfit) and average over their predictions. If models

are independent, the average has the same bias but lower variance than the individual models. Hence, higher total accuracy.

A random forest is a bagging method that makes predictions by combining predictions by several decision trees.

Parameters: m (number of trees), k (number of features to split)

1. Create m bootstrap samples from training data
2. Learn a decision tree for each bootstrap sample
   (a) At each node in the decision tree choose k features at random and choose the best split amond them
   (b) No pruning
3. Combine predictions (majority vote or averaging)

Random forests learn trees without pruning and therefore they are flexible and have low bias. Choosing splits among a random subsets of features forces the trees to be different and makes them less dependent with each other. Thus, the variance is lower.

Grading:

4.1 (4p): Learn several classifiers using bootstrap samples 2p, Explanation of what is a bootstrap sample 1p, Prediction using majority vote or averaging 1p

4.2 (3p): Reduces variance without changing bias 2p, Explaining why the variance is reduced 1p

4.3 (3p): Learn decision trees on bootstrap samples 1.5p, Features for each split are selected randomly 1.5p

Points are deducted for missing details and incorrect claims

5 Machine learning advice

Error/problem	**Diagnostics**	**Fix**
Inappropriate model selection procedure	Training and validation data overlap	Use non-overlapping training and validation sets
Estimate of generalization error is biased	Final model is not tested using unseen data	Test the final model using an unseen test data
Inappropriate performance measure	RMSE is for regression tasks	Use a performance measure that is appropriate for classification problems. For example, accuracy.
Overfitting	From confusion matrices we can infer that training accuracy is 100% while validation accuracy is about 73% (we ignore 20 correctly classified data points that were already in training data). Furthermore, decision boundary is complex and 1-nearest neighbor is known to be prone to overfit.	Use a proper model selection procedure. Additionally, the usual ways to avoid overfitting

Grading:

2.5 points for each of the items above (1.5p for recognizing a problem and 1p for a reasonable solution)

0.5-1 extra points can be given for other insightful observations (max 2 points). (E.g., split training and validation sets randomly)

Incorrect claims lead to deductions of 0.5-1 points each (E.g. claiming that RMSE is computed incorrectly)

Total points are always at least 0 and at most 10.