INF264 Introduction to machine learning
Fall 2019

Exam

4.12.2019

This exam has 5 tasks on 6 pages. You can get at most 50 points.
No aids permitted.

1 Basic concepts (10p)

Give short answers (about one paragraph) to the following questions:

1. What is overfitting and why is it a problem?
2. What is the goal of dimensionality reduction and why is dimensionality reduction useful?
3. Which performance measures are well suitable for imbalanced data in the classification setting? Why?
4. What is information gain and where is it used?
5. Consider linear models (linear regression and logistic regression). How can they be adapted to handle non-linear data?

2 k-means clustering (10p)

You are given a data set $D = \{0, 2, 3, 5, 8\}$ consisting of 5 one-dimensional points. Find a 2-means clustering of D by simulating Lloyd's algorithm with initial cluster centers 2 and 5. Show intermediate steps.

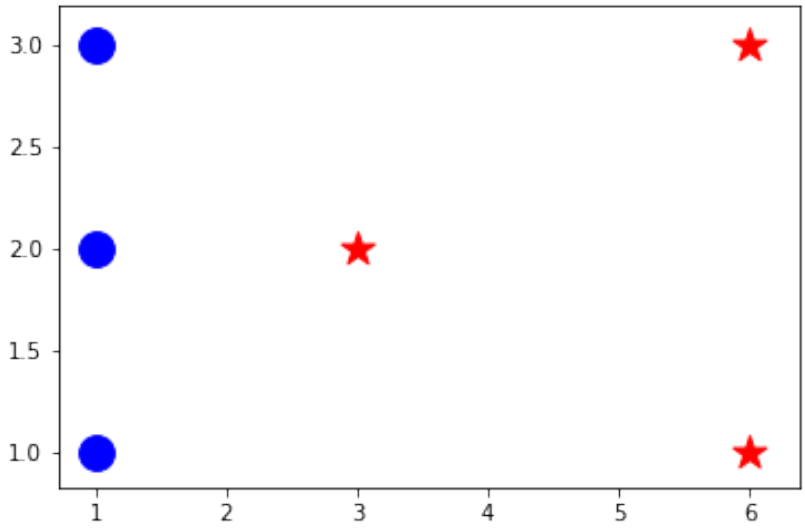

Figure 1: Observations for SVM. A blue circle denotes class -1 and a red star denotes class 1.

3 Support vector machines (10p)

We have observed the following 6 points:

x_1	x_2	y
1	1	-1
1	2	-1
1	3	-1
3	2	1
6	1	1
6	3	1

Data points lie in a 2-dimensional space. The class label is either $y = -1$ or $y = 1$. Figure 1 illustrates the data.

Consider a linear hard margin support vector machine (SVM). Complete the following tasks:

1. Draw the decision boundary. Draw the margins. What are the support vectors? (Note that the points and curves do not have to be exactly correctly drawn. It is enough that we can see that you have understood the idea.)

2. What is the training accuracy of the SVM obtained in the previous step?

3. Evaluate a SVM by conducting 6-fold cross validation on this data set. What is the validation accuracy? (You do not have to explicitly draw all data sets if you can justify the results in another way)

Justify your answers.

4 Bagging and random forests (10p)

Answer the following questions about bagging and random forests:

1. What is bagging?

2. Why can bagging help to increase accuracy? (Hint: you can use the bias-variance tradeoff)

3. How does a random forest differ from a standard bagging model?

5 Machine learning advice (10p)

Task: Read the description of the machine learning pipeline of AwesomeProducts Inc below. Write a short report to help AwesomeProducts Inc to improve their machine learning solution. Specifically, write (a) what are they doing wrong or what problems do they have and (b) how these errors or problems should be fixed.

Description of the machine learning pipeline of AwesomeProducts Inc

After graduating, you have got a job as a machine learning consultant at the Machine Learning Experts Inc. Your first assignment is to help a company called AwesomeProducts Inc. They are developing an app that helps students to optimize their time usage by predicting whether a student will pass an exam given time spend on doing exercises and attending lectures.

Software developers at AwesomeProducts have heard that instead of writing a program by yourself, you can let your computer to learn it. They have

also heard that with easy-to-use libraries like `sklearn`, anybody can do machine learning projects. AwesomeProducts has run a pilot project in machine learning. However, the results have been rather dissappointing. Before abandoning machine learning as a totally useless tool, they have decided to ask an expert opinion.

CTO of AwesomeProducts explains their machine learning pipeline:

> We started by collecting data by interviewing students about their time usage and exam results. In total, we collected **150** samples. Each sample had 2 features: the proportion of exercises completed and the proportion of lectures attended. The label has two possible values: *pass* or *fail*.
>
> To get proper generalization results, we divided the data into training and validation sets so that the first **120** samples were in the training data and the last **50** samples in the validation data.
>
> We decided to use a k-nearest neighbor classifier. We considered hyperparameter values $k = 1, 3, 5, 7, 9$. We measured the performance using root mean squared error (RMSE). That is, we computed
>
> $$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\big(y_i - f(\mathbf{x}_i)\big)^2},$$
>
> where y_i is the label of the ith data point and $f(\mathbf{x}_i)$ is the prediction for the ith data point. To be able to compute RMSE, we encoded the labels *pass* and *fail* with 0 and 1, respectively.
>
> Training and validation RMSEs for different values of k can be seen in Figure 2. We selected the model that gave the smallest validation RMSE. That is, $k = 1$.
>
> We also did some sanity checks. We produced confusion matrices for the chosen classifier; see Figure 3. We were delighted to see that the classifier gave 100% accuracy on training data! Figure 4 illustrates the decision boundary. It is clear that the 1-nearest classifier is very flexible and can model almost any decision boundary.

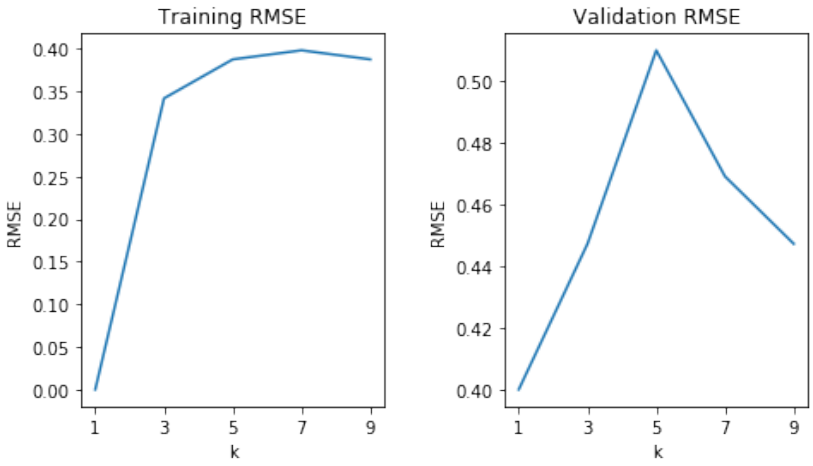

Figure 2: RMSEs on training and validation data

		Prediction	
		fail	*pass*
True	*fail*	54	0
value	*pass*	0	66

		Prediction	
		fail	*pass*
True	*fail*	23	3
value	*pass*	5	19

Figure 3: Confusion matrices for the 1-nearest neighbor classifier on training data (left) and validation data (right).

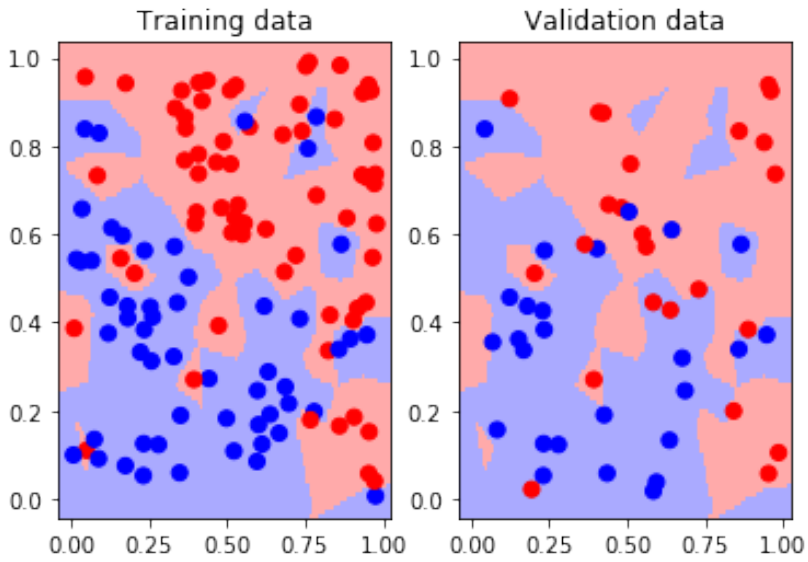

Figure 4: Training and validation data points (*fail* is blue and *pass* is red) and decision boundaries. The background color specifies the prediction of the 1-nearest neighbor classifier.

> Based on the validation error, we expected that the selected 1-nearest neighbor classifier gives RMSE=0.4 on unseen data. However, we tried the model in practice but we got much worse performance. It seems that machine learning does not work after all.
>
> What can we do? I heard that having more data would help but unfortunately students have already left for Christmas holiday so we have to work with the 150 samples that we already have.