Academic year 2019-2020, January 21, 8h30

Machine learning

Clearly write your name on every answer sheet, as well as on all question sheets. Answer each question separately, and clearly mention the number of every question next to your answer. If you don't provide an answer for a question, clearly mention the question number and write "No answer". You can answer either in Dutch or English.
No laptops, calculators, PDAs, phones or Internet access is allowed. Hand in all sheets, including question sheets.

---

1. **Bias-variance trade-off (3pt)**
   Explain what effect will the following operations have on the bias and variance of your model. Fill in one of "increases", "decreases" or "no change" in each of the cells of the table.

	Bias	Variance
Regularizing the weights in a linear regression model		
Increasing k in k-nearest neighbor models		
Pruning a decision tree (to a certain depth for example)		
Increasing the number of hidden nodes in an artificial neural network		
Using dropout to train a deep neural network		
Removing all the non-support vectors in SVM		

2. **Clustering (6pt)**

   (a) Describe two ways to measure the quality of the output of a clustering algorithm and explain how a statistical framework can be used to test whether the result is significantly different from random data.

   (b) Give an algorithmic description of the DBSCAN algorithm.

   (c) Explain the difference between the DBSCAN and KMEANS clustering algorithms concerning their ability to deal with clusters of different shapes.

3. **Support Vector Machines (4pt)**
Given a classification problem where the goal is, given a training set, to predict the test data points as good as possible. Furthermore we know that the data is coming from sensors that are very noisy, so the model should avoid to attach too much attention to specific data points. To solve this problem we will use a Support Vector Machine (SVM) with a quadratic kernel (polynomial kernel of degree two). The dataset is given in Figure 1, and the slack penalty C of the SVM will thus determine how the decision boundary looks like.

(a) Where would be the decision boundary for very big values of C (for example when C goes to infinity)? Draw your answer on the figure below. Motivate why (on your answer sheet).

(b) Indicate on the figure where the decision boundary would be when C goes to 0. Draw your answer on the figure and motivate why.

(c) Draw a datapoint that does not change the decision boundary for very large values of C. Motivate why,

(d) Draw a datapoint that drastically changes the decision boundary for very large values of C. Motivate why.

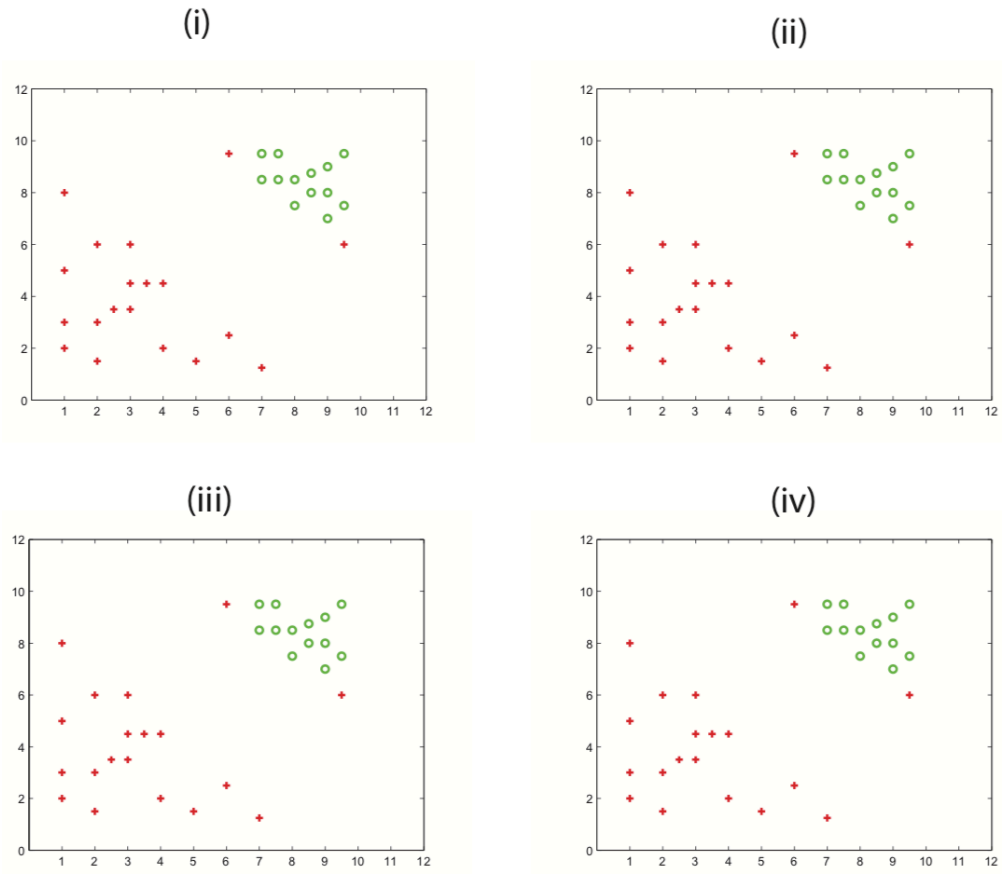

Figure 1: SVM

4. **Bayesian classifiers (1pt)**
   The key assumption of a naive Bayes (NB) classifier is that features are independent, which is not always desirable. Suppose that linear principal components analysis (PCA) is first used to transform the features, and NB is then used to classify data in the low-dimensional space defined by the first few principal components. Is the following statement true? Justify your answers.

   "The independent assumption of NB would now be valid with PCA transformed features because all principal components are orthogonal and hence uncorrelated."

5. **Method design (2pt)**

   (a) Assume we have a set of data from patients who have visited the Ghent University hospital during the year 2019. A set of features (e.g., temperature, height,...) have been also extracted for each patient. Our goal is to decide whether a new visiting patient has any of diabetes, heart disease, or Alzheimer (a patient can have one or more of these diseases). We have decided to use a neural network to solve this problem. We have two choices: either to train a separate neural network for each of the diseases or to train a single neural network with one output neuron for each disease, but with a shared hidden layer. Which method do you prefer? Justify your answer.

   (b) Describe a performance measure that could be used for a machine learning problem where you have only positive and unlabeled data points.

   (c) In which scenario can we apply McNemar's test to compare machine learning models ?

6. **True or False (4 pt)**
Are the following statements True or False ? If True, explain in at most two sentences. If False, explain why or give a counterexample in at most two sentences.

(a) Suppose that $X_1, \ldots, X_m$ are categorical input attributes and Y is a categorical output attribute. Suppose we plan to learn a decision tree without pruning. If X_i and Y are independent in the distribution that generated this dataset, then X_i will not appear in the decision tree.

(b) When you would run K-means on a dataset with overlapping clusters, the cluster centers found by K-means would be pushed further apart compared to the true cluster centers.

(c) In the context of classification, it is always advantageous to remove features with a very high variance.

(d) A classification model with 100% accuracy on the training set and 70% accuracy on the test set is better than a model with 70% accuracy on the training set and 75% accuracy on the test set.

(e) PCA is a supervised form of dimensionality reduction.

(f) If X and Y are independent and $X > 1$, then $\text{Var}[X + 2Y^2]=\text{Var}[X]+4\text{Var}[Y^2]$ and $\text{E}[X^2 - X] \geq \text{Var}[X]$

(g) The following is a good procedure for performing feature selection. A project team performed a feature selection procedure on the full data and reduced their large feature set to a smaller set. Then they split the data into test and training portions. They built their model on training data using several different model settings, and report the best test error they achieved.

(h) A multi-layer perceptron model (MLP) with $N > 1$ hidden layers and linear activation functions can always be represented as e neural network without hidden layer.