KTH Computer Science and Communication

Exam in DD2431 Machine Learning
2017-03-18, kl 9.00 – 13.00

Aids allowed: *calculator*, *language dictionary*.

In order to pass, you have to fulfill the requirements defined both in the A- and in the B-section.

A Questions on essential concepts

Note: As a prerequisite for passing you must give the correct answer to almost *all* questions. Only *one* error will be accepted, so pay good attention here.

A-1 Probabilistic Learning

What is the goal of *maximum a posteriori* estimation?

Find the model parameters that:

a) maximize the prior.

b) maximize a convex optimality criterion.

c) optimize the likelihood of the new observations in conjunction with a priori information.

A-2 Naive Bayes Classifier

What is the underlying assumption unique to a *naive Bayes* classifier?

a) All features are regarded as conditionally independent.

b) The number of features (the dimension of feature space) is large.

c) A Gaussian distribution is assumed for the feature values.

A-3 Shannon Entropy

Consider a single toss of fair coin. Regarding the uncertainty of the outcome {head, tail}:

a) The entropy is equal to one bit.

b) The entropy is equal to two bits.

c) The entropy is not related to uncertainty.

A-4 Regression

In regression, regularization is a process of introducing additional term, so-called shrinkage penalty. Which one of the three methods includes the additional term.

a) Least squares.

b) Ridge regression.

c) k-NN regression.

A-5 Artificial Neural Networks

What happens *during training* in an artificial neural network?

a) Weights are adjusted to minimize the output error.

b) Training samples are sorted according to their likelihood.

c) Nodes are added to maximize the information gain.

A-6 Support Vector Machine

What is the consequence of using a *kernel*-function in a *support vector machine*?

a) Sample averages can be computed very efficiently.

b) Classification takes place in a virtual high-dimensional space.

c) Overlapping distributions will still give stable results.

A-7 Ensemble Learning

Which one below best describes the characteristics of Ensemble methods in machine learning?

a) Weak models are trained and combined.

b) Ensemble methods are aimed to deal with the curse of dimensionality.

c) The performance of ensemble learning is proportional to the number of models combined.

A-8 PCA

What is the main role of the principal component analysis (PCA) in the *Subspace Method* for classification?

a) To compute a subspace that represents the training data distribution in each class.

b) To choose a single representative sample in the training data in each class.

c) To compute the borderline samples to define the separation surface between classes.

Note: Your answers need be on a separate solution sheet (**we will not receive this page**).

B Graded problems

A pass is guaranteed with the required points for 'E' below (excluding bonus) in this section *and* the prerequisite in the A-section.

Preliminary number of points required for different grades:

$$
\begin{aligned}
22 \leq p \leq 24 &\rightarrow A \\
19 \leq p < 22 &\rightarrow B \\
16 \leq p < 19 &\rightarrow C \\
13 \leq p < 16 &\rightarrow D \\
8 \leq p < 13 &\rightarrow E \\
0 \leq p < 8 &\rightarrow F
\end{aligned}
$$

B-1 Terminology (4p)

For each term (a–h) in the left list, find the explanation from the right list which *best* describes how the term is used in machine learning.

a) Random Forests

b) RANSAC

c) Dropout

d) k-means

e) Curse of dimensionality

f) Gini impurity

g) Expectation Maximization

h) Projection length

1) The length of cast shadow

2) An approach to train artificial neural networks

3) Random strategy for amplitude compensation

4) Robust method to fit a model to data with outliers

5) Ensemble of decision trees

6) Probability at a later time

7) Method for estimating the mean of k observations

8) Clustering method based on centroids

9) Convex optimization

10) Issues in data sparsity in space

11) A definition of predictability

12) A measure of inequality between samples

13) Implementation of the bag-of-words model

14) Similarity measure in the subspace method

15) Algorithm to learn with latent variables

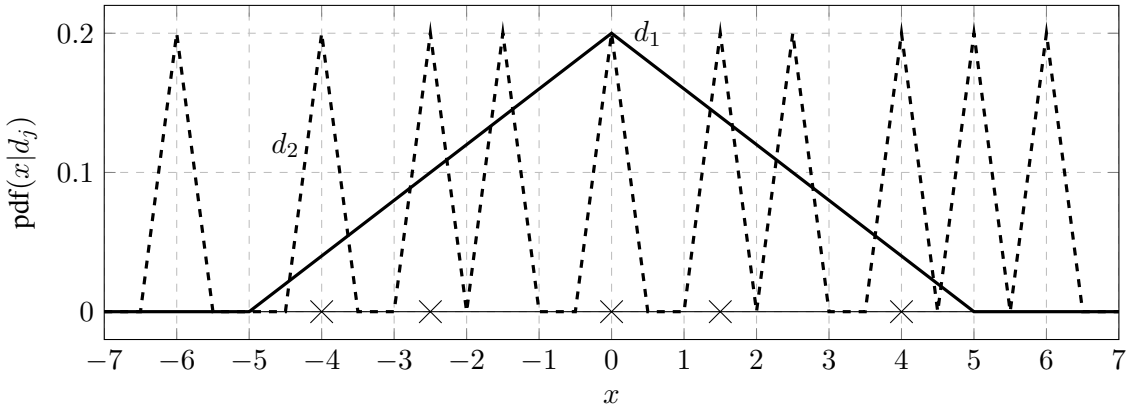

Figure 1. Illustration for Problem B-3

B-2 Probability based learning (3p)

Suppose you need to design an identity verification system based on face recognition whose goal is to confirm or reject the identity claimed by each user. The system is only supposed to work with a close set of $N > 1$ individuals. Users are assumed to claim any of the N identities uniformly at random. Call α the probability of false acceptance and β the probability of false rejection of the system.

a) What is the a priori probability that the claimed identity is correct?

b) What are the conditions on α and β to make sure that the claimed identity is more likely to be correct if the system accepts the user and more likely to be incorrect if the system rejects the user?

c) What are the conditions on α and β from the previous point if you assume equal error rates?

B-3 Probability based Learning (3p)

Figure 1 shows five data points on the real number line (x-axis) denoted by the symbol $\times$ and two probability distribution functions (PDFs) denoted by d_1 (continuous line) and d_2 (dashed line).

a) Which PDF fits the data best according to the Likelihood criterion, assuming that the data is i.i.d. (independent and identically distributed)?

b) Assuming the shapes are all triangular and symmetric around the center, how many parameters do you need to define d_1 and d_2 respectively? **Note:** d_2 can be considered as a mixture of distributions in the form of d_1.

c) What are the risks of comparing model fit using the likelihood as in point **a**). Explain referring to this particular case.

B-4 Classification (2p)

Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use 1-nearest neighbor and get an average error rate (averaged over both test and training data sets) of 20%. Next we use the Discriminant function and get an error rate of 18% on the training data. We also get the average error rate (averaged over both test and training data sets) of 24%.

a) What was the error rate with the Discriminant function on the test set?

b) Based on these results, indicate the method which we should prefer to use for classification of new observations, with a simple reasoning.

B-5 Information gain (3p)

You have booked a flight for tomorrow, but now there are some risks that it might be cancelled due to two factors: typhoon and strike. Your estimate on flight cancellation due to the weather, i.e. typhoon, is 20%. Independently, the probability of cancellation due to strike is 50%.

a) What is the probability that there will be a flight cancellation due to one or both of the factors?

b) How unpredictable is it that the flight is cancelled, either due to the weather or the strike (or both)? Answer in terms of Entropy, measured in bits.

c) You realized that you can find out on the airport website tomorrow morning if there will be a strike or not (which we can assume reliable). What is the expected information gain from checking it on the website?

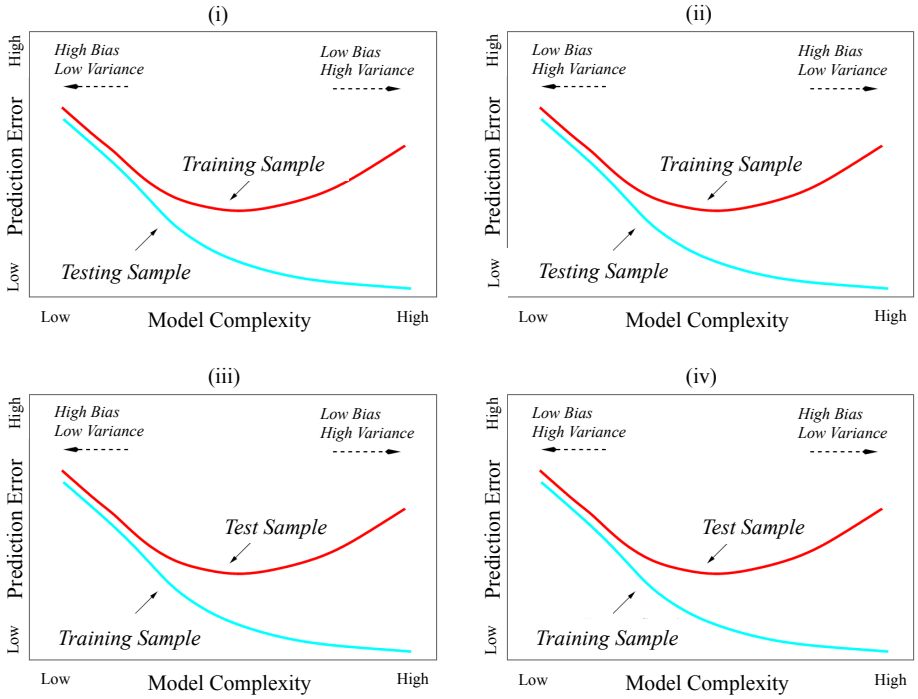

Figure 2. Typical behavior of prediction error plotted against model complexity.

B-6 Bias and Variance (3p)

a) One of the four subfigures (i)-(iv) in Figure 2 displays the typical trend of prediction error of a model for training and testing data with comments on its bias and variance, {high, low}. Which one of the four figures most well represents the general situation?

b) Briefly explain the main reason why the prediction errors have different trend for training samples and test samples.

c) Now consider the specific case of using *Bagging* by an ensemble of decision tree classifiers. What sort of improvememt can be expected in the ensemble predictions in terms of *bias* or *variance* of the classifier as a whole?

B-7 Support Vector Machines (3p)

In the SVM-lab, you used a predefined procedure (`qp`) for finding the solution to a quadratic problem on standard form. We therefore reformulated the SVM task

$$\text{minimize } \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j t_i t_j \mathcal{K}(\vec{x}_i, \vec{x}_j) - \sum_i \alpha_i \qquad \text{while } \alpha_i \geq 0 \quad \forall i$$

by rewriting it into matrix form to match the parameters of the procedure:

$$\frac{1}{2}\vec{\alpha}^T P \vec{\alpha} - \vec{\alpha} \cdot \vec{1} \qquad \text{where } \vec{\alpha} \geq \vec{0}$$

When debugging the code it is often useful to calculate the parameters by hand. What would be the correct contents of the matrix P if we use the polynomial kernel with exponent $p = 3$ and training data consisting of these three samples

$$x_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \qquad x_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \qquad x_3 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

where x_1 and x_2 are positive samples, and x_3 a negative sample?

B-8 Perceptron Learning (3p)

Use the *perceptron learning rule* to test if the following dataset is linearly separable.

Positive samples	Negative samples
[0 0 1 0 0]	[1 1 0 1 1]
[1 0 1 1 0]	[0 1 0 0 1]
[1 1 0 0 1]	[0 0 1 1 0]
[1 0 0 0 0]	[0 1 1 1 1]