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.

Solution: c

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.

Solution: a

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.

Solution: a

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.

Solution: b

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.

Solution: a

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.

Solution: b

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.

Solution: a

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.

Solution: a

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

Solution: a-5, b-4, c-2, d-8, e-10, f-11, g-15, h-14

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?

Solution:

a) We call C the event that the claimed identity matches the true identity, $\bar{C}$ is the negation (mismatching identities). We also call V the event that the face recognition system will accepts the identity, and $\bar{V}$ that it will reject it. A priori, given N individuals, the probability of the claimed identity being correct is $P(C) = \frac{1}{N}$ and $P(\bar{C}) = 1 - P(C) = \frac{N-1}{N}$.

b) The probability of false acceptance and false rejection are the probability of the wrong decision given the true identity, that is:

$$\begin{aligned} P(V|\bar{C}) &= \alpha, \quad \text{and consequently, } P(\bar{V}|\bar{C}) = 1 - \alpha \\ P(\bar{V}|C) &= \beta, \quad \text{and consequently, } P(V|C) = 1 - \beta \end{aligned}$$

We want to ensure that the posterior probability of the claimed identity is highest for the corresponding decision of the verification system, that is

$$\begin{aligned} P(C|V) &> P(\bar{C}|V), \quad \text{and} \\ P(\bar{C}|\bar{V}) &> P(C|\bar{V}) \end{aligned}$$

using Bayes rule (and simplifying the common denominator) the two inequalities become:

$$\begin{aligned} P(V|C)P(C) &> P(V|\bar{C})P(\bar{C}), \quad \text{and} \\ P(\bar{V}|\bar{C})P(\bar{C}) &> P(\bar{V}|C)P(C) \end{aligned}$$

Substituting α, β and N, and rearranging:

$$\begin{aligned} (N-1)\alpha + \beta &< 1, \quad \text{and} \\ (N-1)\alpha + \beta &< N - 1 \end{aligned}$$

Because we have assumed $N \in \mathbb{N}$ and $N > 1$, the second inequality is always verified if the first is, so the final answer is:

$$(N-1)\alpha + \beta < 1$$

and of course α and β must be bounded between 0 and 1 because they are probabilities.

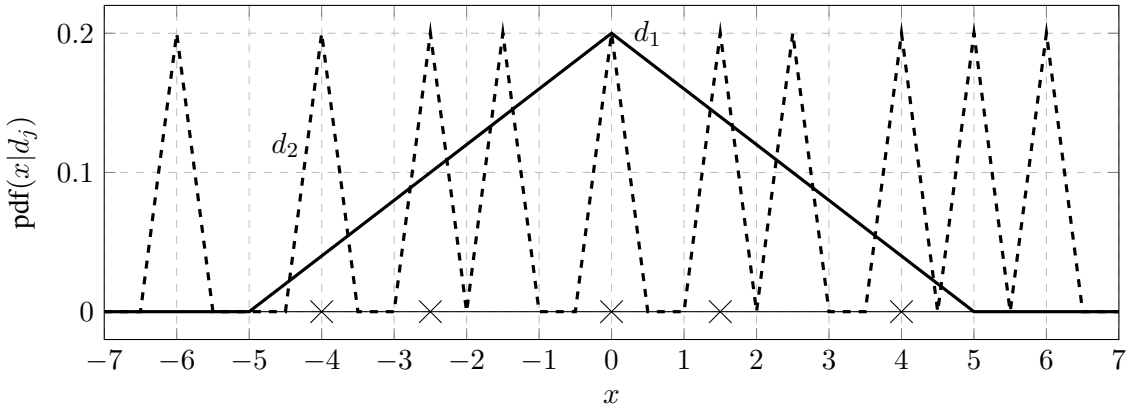

Figure 1. Illustration for Problem B-3

c) if we assume equal error rates, then $\alpha = \beta$ and the above inequality becomes:

$$\alpha = \beta < \frac{1}{N}, \quad \alpha, \beta \geq 0.$$

Intuitively, if the number of user increases, in order to ensure the correct functionality of the system, we have to impose much more demanding requirements on the discriminative power of the methods (that is lower 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.

Solution:

a) The likelihood $\mathcal{L}(\mathcal{D})$ is an indication of how well a probability distribution (model) fits the data. If the data points are i.i.d. (independent and identically distributed), the likelihood of the data can be expressed as the product of the likelihood of the single points:

$$\mathcal{L}_j(\mathcal{D}) = \prod_i \text{pdf}(x_i|d_j)$$

In the case in the figure, it is evident that d_2 (dashed distribution) has a higher likelihood, because for each point x_i it is always true that $\text{pdf}(x_i|d_2) \geq \text{pdf}(x_i|d_1)$. If we want to express this numerically, we have, for d_2:

$$\mathcal{L}_2(\mathcal{D}) = \prod_i \text{pdf}(x_i|d_1) = 0.2^5 = 0.00032$$

And for d_2, after some simple geometry:

$$\mathcal{L}_1(\mathcal{D}) = \prod_i \text{pdf}(x_i|d_1) = \frac{1}{5}0.2\times\frac{2.5}{5}0.2\times0.2\times\frac{3.5}{5}0.2\times\frac{1}{5}0.2 = 0.2^5\frac{7}{2^2\times5^4} = 0.0000045$$

b) To define d_1 we need just two parameters: the center of the distribution μ (that also coincides with the expected value because of the symmetry), and a parameter related to the spread. For simplicity the second parameter could be the value of the PDF at the center: $\delta = \text{pdf}(\mu|d_1)$. Because the area below the distribution must be equal to 1, the distribution is fully defined if we know μ and δ. In the particular example in the figure, we have $\mu = 0$ and $\delta = 0.2$.

If we consider d_2 as a mixture of N distributions in the form of d_1, we need $N\times2$ parameters (from d_1), plus N weights. Because the weights must sum to 1, only $N-1$ weights are free parameters. In total we have $2\times N + N - 1 = 3\times N - 1$ parameters. From the figure we see that d_2 is comprised by 10 components, so the number of parameters would be 29 at most. From the figure we could also make some simplifying assumptions: all the weights are equal and therefore equal to 0.1, all the δ parameters are equal (in the example they will be equal to 2.0 considering the weights). In this special case, the only free parameters would be the 10 μ_is.

c) When comparing models of different complexity, the likelihood can be made indefinitely high by overfitting. In the specific example, d_2 fits very well the specific points, but is not likely to fit new points if the underlying distribution from which they are drawn is not exactly equal to d_2. On the other hand, d_1 has a lower fit for this specific points, but will have a similar fit for new points as well.

The following discussion is outside the scope of the question and is included for completeness: Better ways of comparing models would involve estimating the posterior of the model given the data by integrating over all possible values of the parameters. This way, more complex models will have a much better fit to the data but only for a much more restricted domain of the parameter space, and therefore, more simple models have a chance to win the comparison. Approximate methods, such as the Bayes Information Criterion, avoid computing a complex integral of the posterior by weighting the likelihood with the number of parameters in the model and the number of data points used for the evaluation.

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.

Solution:

a) 30%

b) Discriminant function because it achieves lower error rate on the test data (30% < 40%).

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?

Solution:

1. 1 - (The probability that there will be no flight cancelletion = 0.80 $\times$ 0.50) = 0.60.
2. -0.6 log_2 0.6 - 0.4 log_2 0.4 $\approx$ 0.971 (bits)
3. With proobability 0.5 you will know for sure (Ent = 0) that the flight is canceled. Otherwise, the only uncertainly that remains is the one caused by the weather which has the entropy: -0.8 log_2 0.8 - 0.2 log_2 0.2 $\approx$ 0.722.
   Expected gain = 0.971 - (0.5 $\times$ 0 + 0.5 $\times$ 0.722) = 0.610 (bits)

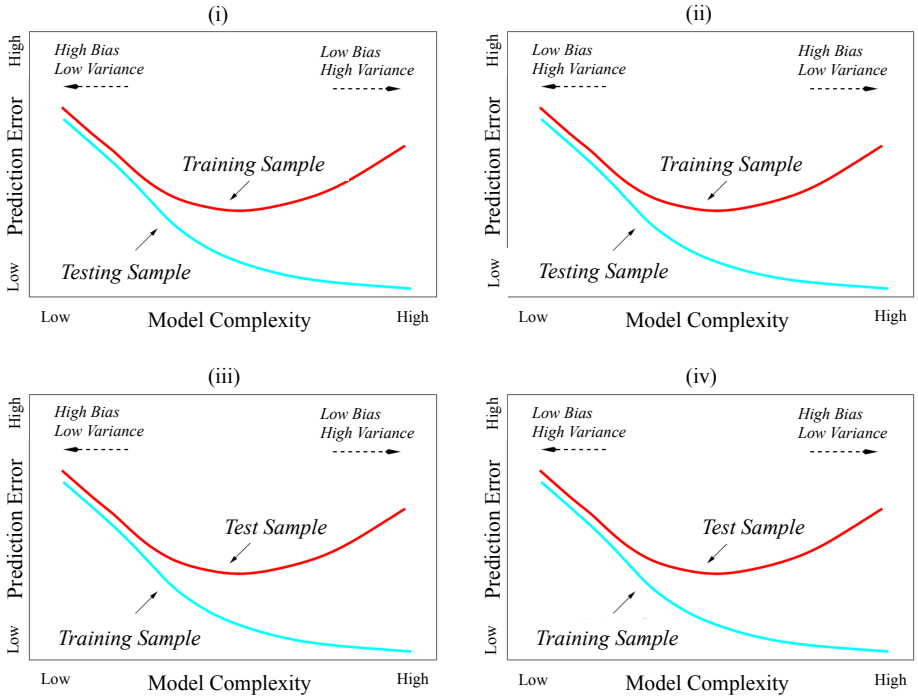

Figure 2. Typical behavior of prediction error plotted against model complexity.

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?

Solution:

a) (iii)

b) Overfitting.

c) Reduction of the variance.

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?

Solution: The kernel to use is

$$\mathcal{K}(\vec{x}, \vec{y}) = (\vec{x}^T \vec{y} + 1)^3$$

$$P = \begin{bmatrix} t_1 t_1 \mathcal{K}(\vec{x}_1, \vec{x}_1) & t_1 t_2 \mathcal{K}(\vec{x}_1, \vec{x}_2) & t_1 t_3 \mathcal{K}(\vec{x}_1, \vec{x}_3) \\ t_2 t_1 \mathcal{K}(\vec{x}_2, \vec{x}_1) & t_2 t_2 \mathcal{K}(\vec{x}_2, \vec{x}_2) & t_2 t_3 \mathcal{K}(\vec{x}_2, \vec{x}_3) \\ t_3 t_1 \mathcal{K}(\vec{x}_3, \vec{x}_1) & t_3 t_2 \mathcal{K}(\vec{x}_3, \vec{x}_2) & t_3 t_3 \mathcal{K}(\vec{x}_3, \vec{x}_3) \end{bmatrix}$$

which gives us

$$P = \begin{bmatrix} 27 & 8 & -8 \\ 8 & 27 & -8 \\ -8 & -8 & 27 \end{bmatrix}$$

B-8 Perceptron Learning (3p)

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

Positive samples	Negative samples
$[0 \quad 0 \quad 1 \quad 0 \quad 0]$	$[1 \quad 1 \quad 0 \quad 1 \quad 1]$
$[1 \quad 0 \quad 1 \quad 1 \quad 0]$	$[0 \quad 1 \quad 0 \quad 0 \quad 1]$
$[1 \quad 1 \quad 0 \quad 0 \quad 1]$	$[0 \quad 0 \quad 1 \quad 1 \quad 0]$
$[1 \quad 0 \quad 0 \quad 0 \quad 0]$	$[0 \quad 1 \quad 1 \quad 1 \quad 1]$

Solution: The perceptron learning rule *always* converges if the dataset is linearly separable. We therefore test by repeatedly applying the rule to the samples.

Choose an arbitrary starting weight vector, e.g. $\vec{w} = [0, 0, 0, 0, 0, 0]$. Note the extra weight (w_0) which is weighing in the constant bias value (here taken to be 1 as usual).

This is how the weights will change as our samples are processed:

Weights						x_0	Sample					Scalar prod.	Result
0	0	0	0	0	0	1	0	0	1	0	0	$0 \leq 0$	Wrong, add sample
1	0	0	1	0	0	1	1	1	0	1	1	$1 > 0$	Wrong, subtract sample
0	-1	-1	1	-1	-1	1	1	0	1	1	0	$-1 \leq 0$	Wrong, add sample
1	0	-1	2	0	-1	1	0	1	0	0	1	$-1 \leq 0$	Ok
1	0	-1	2	0	-1	1	1	1	0	0	1	$-1 \leq 0$	Wrong, add sample
2	1	0	2	0	0	1	0	0	1	1	0	$4 > 0$	Wrong, subtract sample
1	1	0	1	-1	0	1	1	0	0	0	0	$2 > 0$	Ok
1	1	0	1	-1	0	1	0	1	1	1	1	$1 > 0$	Wrong, subtract sample
0	1	-1	0	-2	-1	1	0	0	1	0	0	$0 \leq 0$	Wrong, add sample
1	1	-1	1	-2	-1	1	1	1	0	1	1	$-2 \leq 0$	Ok
1	1	-1	1	-2	-1	1	1	0	1	1	0	$1 > 0$	Ok
1	1	-1	1	-2	-1	1	0	1	0	0	1	$-1 \leq 0$	Ok
1	1	-1	1	-2	-1	1	1	1	0	0	1	$0 \leq 0$	Wrong, add sample
2	2	0	1	-2	0	1	0	0	1	1	0	$1 > 0$	Wrong, subtract sample
1	2	0	0	-3	0	1	1	0	0	0	0	$3 > 0$	Ok
1	2	0	0	-3	0	1	0	1	1	1	1	$-2 \leq 0$	Ok
1	2	0	0	-3	0	1	0	0	1	0	0	$1 > 0$	Ok
1	2	0	0	-3	0	1	1	1	0	1	1	$0 \leq 0$	Ok
1	2	0	0	-3	0	1	1	0	1	1	0	$0 \leq 0$	Wrong, add sample
2	3	0	1	-2	0	1	0	1	0	0	1	$2 > 0$	Wrong, subtract sample
1	3	-1	1	-2	-1	1	1	1	0	0	1	$2 > 0$	Ok
1	3	-1	1	-2	-1	1	0	0	1	1	0	$0 \leq 0$	Ok
1	3	-1	1	-2	-1	1	1	0	0	0	0	$4 > 0$	Ok
1	3	-1	1	-2	-1	1	0	1	1	1	1	$-2 \leq 0$	Ok
1	3	-1	1	-2	-1	1	0	0	1	0	0	$2 > 0$	Ok
1	3	-1	1	-2	-1	1	1	1	0	1	1	$0 \leq 0$	Ok
1	3	-1	1	-2	-1	1	1	0	1	1	0	$3 > 0$	Ok
1	3	-1	1	-2	-1	1	0	1	0	0	1	$-1 \leq 0$	Ok
1	3	-1	1	-2	-1	1	1	1	0	0	1	$2 > 0$	Ok
1	3	-1	1	-2	-1	1	0	0	1	1	0	$0 \leq 0$	Ok
1	3	-1	1	-2	-1	1	1	0	0	0	0	$4 > 0$	Ok
1	3	-1	1	-2	-1	1	0	1	1	1	1	$-2 \leq 0$	Ok

The weight vector $[1, 3, -1, 1, -2, -1]$ classifies all the samples correctly. We can therefore draw the conclusion that the samples *are* linearly separable.