KTH Computer Science and Communication

Exam in DD2421 Machine Learning
2018-03-16, kl 14.00 – 18.00

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 choose the correct answer to almost *all* questions. Only *one* error will be accepted, so pay good attention here.

A-1 Probabilistic Learning

Suppose that for two events A and B we can write that $P(A|B) = P(A)$. This means that:

a) A is more likely than B

b) A is deterministic

c) A and B are independent

Solution: c

A-2 Naive Bayes Classifier

What is the main assumption in the Naive Bayes Classifier?

a) Features are independent from the classes.

b) The classes are independent.

c) Features are conditionally independent.

Solution: c

A-3 Shannon Entropy

Consider a single toss of *skewed* coin (it is likely to show one side more than the other side). Regarding the uncertainty of the outcome {head, tail}, the entropy

a) is equal to one bit.

b) is smaller than one bit.

c) does not explain the uncertainty.

Solution: b

A-4 Regression and Classification

Choose the most proper statement reflecting the output formats of regression and classification.

a) Discrete for regression and continuous-valued for classification.

b) Discrete for classification and continuous-valued for regression.

c) They are both continuous-valued.

Solution: b

A-5 Artificial Neural Networks

What values of an artificial neural network are adjusted during training when using the Back-Propagation algorithm?

a) Weights and thresholds

b) Means and covariance

c) Labels of training samples

Solution: a

A-6 Kernels

What role does the *kernel function* have in a support vector machine?

a) It computes the dot-product in a high-dimensional space.

b) It integrates the error over the whole data set.

c) It updates the weights in the network.

Solution: a

A-7 Ensemble Learning

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

a) Ensemble learning is not well-suited to parallel computing.

b) Diverse models are trained and combined.

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

Solution: b

A-8 Principal Component Analysis (PCA)

Which one is considered as the main purpose of the principal component analysis (PCA)?

a) To find the least squares fit.

b) To reduce the effective number of variables.

c) To use class labels in an optimal way.

Solution: b

Note: Your answers (eight of them) need be on a solution sheet (**this page will not be received**).

B Graded problems

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

Preliminary number of points required for different grades:

$$
\begin{aligned}
24 \leq p \leq 27 &\rightarrow A \\
20 \leq p < 24 &\rightarrow B \\
16 \leq p < 20 &\rightarrow C \\
12 \leq p < 16 &\rightarrow D \\
9 \leq p < 12 &\rightarrow E \\
0 \leq p < 9 &\rightarrow F
\end{aligned}
$$

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) Occams's razor

b) RANSAC

c) Dropout

d) Fisher's criterion

e) Support Vector

f) Perceptron Learning

g) k-means

h) Posterior probability

1) Clustering method based on centroids

2) A concept of accepting high model complexity

3) Method to find separating hyperplanes

4) Conditional probability taking into account the evidence

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

6) Sudden drop of performance

7) Random strategy for amplitude compensation

8) Learning trying to mimic human vision

9) Data point affecting the decision boundary

10) A principle to choose the simplest explanation

11) An approach to train artificial neural networks

12) Probability at a later time

13) Vector representation of a feature

14) Method for estimating the average of k observations

15) An approach to find useful dimension for classification

Solution: a-10, b-5, c-11, d-15, e-9, f-3, g-1, h-4 (f-11 is also allowed.)

B-2 Maximum a Posteriori Classifier (3p)

We use a maximum a posteriori classifier to distinguish two classes depending on the value of a single feature $x \in \mathbb{R}$. Each class-conditional probability distribution function is assumed to be Gaussian:

$$\text{pdf}(x|c_i) \quad = \quad \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right),$$

with $i \in \{0, 1\}$. We assume that the classes are equally likely a priori. We also assume that the means of the two classes are opposite with respect to the origin, and that the variance is the same for both class-conditional distributions:

$$\mu_0 = -\mu, \qquad \sigma_0^2 = \sigma^2,$$
$$\mu_1 = \mu, \qquad \sigma_1^2 = \sigma^2.$$

a) what is the form of $P(c_1|x)$, the posterior probability distribution for class 1, as a function of x, μ and σ? Simplify as much as you can.

b) what is the range of values for x that are assigned to class 1? Does this range depend on μ and/or σ?

c) what is the range of values for x for which the posterior of class 0 is above 0.9? Does this range depend on μ and/or σ?

Hint: if you don't have a calculator, it may help to know that $e^{2.2} \approx 9$.

Solution:

a) The posterior, according to Bayes rule is:

$$\begin{aligned}
P(c_1|x) &= \frac{\text{pdf}(x|c_1)P(c_1)}{\text{pdf}(x|c_0)P(c_0) + \text{pdf}(x|c_1)P(c_1)} = \ldots \text{imposing equal priors} \ldots \\
&= \frac{\text{pdf}(x|c_1)}{\text{pdf}(x|c_0) + \text{pdf}(x|c_1)} = \frac{1}{1 + \frac{\text{pdf}(x|c_0)}{\text{pdf}(x|c_1)}} = \ldots \text{substituting our Gaussians} \ldots \\
&= \frac{1}{1 + \frac{\exp\left(-\frac{(x+\mu)^2}{2\sigma^2}\right)}{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}} = \frac{1}{1 + \exp\left(-\frac{(x+\mu)^2}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^2}\right)} = \\
&\quad \ldots \text{expanding and simplifying} \ldots \\
&= \frac{1}{1 + \exp\left(-\frac{2\mu}{\sigma^2}x\right)}.
\end{aligned}$$

Note that this is a sigmoid function:

$$P(c_1|x) \quad = \quad \text{sig}(wx),$$

where w is the ratio between the distance 2μ between the means, and the variance of each class σ^2.

b) we choose class 1 iff $P(c_1|x) > \frac{1}{2}$, which is true if $\exp\left(-\frac{2\mu}{\sigma^2}x\right) < 1$. The exponential function is equal to 1 when its argument is equal to zero, and increases with the argument. The range of values for which the above is true is, therefore $-\frac{2\mu}{\sigma^2}x < 0$, or, equivalently, $x > 0$ (if we assume $\mu > 0$). This is intuitive because the problem is symmetric around the origin. This solution only depends on the sign of μ and not on the value of the parameters μ and σ^2.

c) because $P(c_0|x) + P(c_1|x) = 1$, the posterior for class 0 is greater than 0.9 iff $P(c_1|x) < 0.1 = \frac{1}{10} = \frac{1}{1+9} \approx \frac{1}{1+\exp(2.2)}$. In order for this to be verified, it must be $\exp\left(-\frac{2\mu}{\sigma^2}x\right) > \exp(2.2)$ which is true iff $-\frac{2\mu}{\sigma^2}x > 2.2$ or

$$x < -2.2\frac{\sigma^2}{2\mu}$$

Note how the higher the variance σ^2 with respect to the mean difference 2μ, the more the range is shifted to the left meaning that the sigmoid has a slow rising slope around the origin. This corresponds to classes that are not well separated. On the contrary, the higher the value of 2μ with respect to σ^2, the closer the range will come to zero, meaning that the sigmoid has a very steep rise around the origin. This corresponds to well separated classes.

B-3 Probability based Learning (3p)

We want to estimate the probability of getting "head" when tossing a possibly biased coin. We perform four experiments (trials). In each trial we toss the coin ten times. The results are shown in the table:

trial	outcomes									
1	1	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	1	1	1	0	0
3	0	0	0	1	0	0	1	0	1	0
4	0	0	0	0	0	1	1	0	0	0

where 1 corresponds to "head" and 0 to "tail".

a) using maximum likelihood, estimate the probability of obtaining "head" in each of the four trials.

b) assuming we performed a trial before the current trials where we estimated the probability of "head" to be 0.2 by tossing the coin 90 times. How can we update the original estimate based on each of the current trials?

c) assume you were given the four estimates from point (a) and the four estimates from point (b), but you were not told how these were obtained. Suggest a way to measure if the estimates in (a) are more or less reliable than those in (b).

Solution:

a) The probability distribution of a binary variable is called Bernoulli and is specified by a single parameter $\lambda = P(1)$, which corresponds in our case to the probability of "head". (the probability of "tail" is simply $P(0) = 1 - \lambda$.) The maximum likelihood estimation of the parameter λ corresponds to the frequency of the outcome 1 in the data, that is:

trial	λ_{ML}
1	1/10 = 0.1
2	4/10 = 0.4
3	3/10 = 0.3
4	2/10 = 0.2

b) we call $N = 90$ the number of tosses and n the number of heads in the original trial. Similarly we call $M = 10$ the number of tosses and m the number of heads in the following trial. Considering the combination of tosses in the original and the new trial, the ML estimate is:

$$\lambda_{\text{ML}}^{\text{all}} = \frac{n+m}{N+M}$$

We can calculate n from the original estimate: $\lambda_{\text{ML}}^{\text{orig}} = \frac{n}{N} = \frac{n}{90} = 0.2$, that gives: $n = 18$. For each of the new trials we have:

trial	$\lambda_{\text{ML}}^{\text{all}}$
1	(18+1)/(90+10) = 0.19
2	(18+4)/100 = 0.22
3	(18+3)/100 = 0.21
4	(18+2)/100 = 0.2

c) a measure of reliability of repeated estimates is the variance of the estimates. If repeated estimates have low variance we can trust each of them more. It is easy to see that the estimates in (b) have lower variance because they are based on a larger number of observations.

B-4 Classification (3p)

Suppose that we take a data set, divide it into training and test sets, and then try out two different classification procedures. We use half of the data for training, and the remaining half for testing. First we use k-nearest neighbor (where $k = 1$) and get an average error rate (averaged over both test and training data sets) of 7%. Next we use Logistic Regression and get an error rate of 8% on the training data. We also get the average error rate (averaged over both test and training data sets) of 10%.

a) What was the error rate with 1-nearest neighbor on the test set?

b) What was the error rate with the Logistic Regression on the test set?

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

Solution:

a) 14%. (Training error for 1-NN is always zero, and therefore the testing error is 14%.)

b) 12%.

c) Logistic Regression because it achieves lower error rate on the test data (12% < 14%).

B-5 Regression with regularization: LASSO (3p)

For a set of N training samples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, each consisting of input vector $\mathbf{x}$ and output y, suppose we estimate the regression coefficients $\mathbf{w}^\top (\in \mathbf{R}^d) = \{w_1, \ldots, w_d\}$ in a linear regression model by minimizing

$$\sum_{n=1}^{N} (y_n - \mathbf{w}^\top \mathbf{x}_n)^2 \quad \text{subject to} \quad \sum_{i=1}^{d} |w_i| \leq s$$

for a particular value of s.

For parts **a**) through **c**), indicate which of {i,ii,iii,iv,v} is correct. *Briefly justify your answer.*

a) As we *increase* s from 0, the *training* error (residual sum of squares, RSS) will:
i. Remain constant.
ii. Steadily increase.
iii. Steadily decrease.
iv. Increase initially, and then eventually start decreasing in an inverted U shape.
v. Decrease initially, and then eventually start increasing in a U shape.

b) Repeat **a**) for *test* RSS.

c) Repeat **a**) for variance and (squared) bias, respectively.

Solution: When $s = 0$, all w_i are zero; the model is extremely simple, predicting a constant and has no variance with the prediction being far from actual value, thus with high bias. As we increase s, all w_i increase from zero toward their least square estimate values while steadily decreasing the *training* error to the ordinary least square RSS and also decreasing bias as the model continues to better fit training data. The values of w_i then become more dependent on training data, thus increasing the variance. The *test* error initially decreases as well, but eventually start increasing due to overfitting to the training data.

a) iii.

b) v.

c) ii for variance, iii for bias.

B-6 Ensemble Methods (3p)

Briefly answer the following questions regarding ensemble methods of classification.

a) What are the two kinds of randomness involved in the design of Random Forests?

b) In Adaboost algorithm, each training sample is given a weight and it is updated according to some factors through an iteration of training weak classifiers. What are the two most dominant factors in updating the weights?

c) In Adaboost algorithm, how are the two factors mentioned in **b)** used?

Solution:

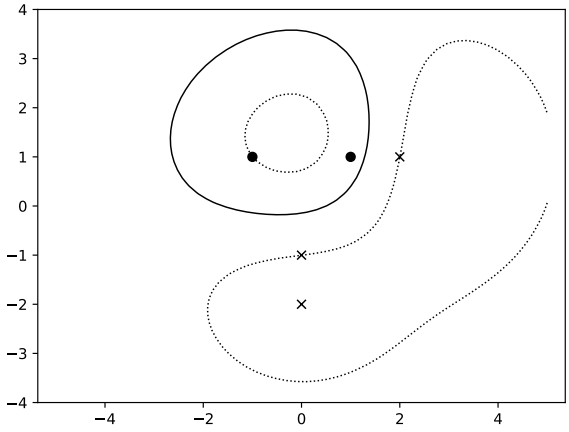
a) Randomness (i) in generating bootstrap replicas and (ii) in possible features to use at each node.

b) The update is according to (i) if the sample was misclassified, and (ii) the reliability of the weak classifier based on the training error; the smaller the training error, the greater the reliability.

c) The weight is increased if misclassified, and decreased if classified correctly.
The reliability is then used as the coefficient.

B-7 Support Vector Machines (3p)

Training a support vector machine has resulted in the decision boundary (solid line) and margin (dotted lines) shown in the figure. Training was done with a RBF (Radial Basis Function) kernel and the C-value was set to 4.0 to allow for slack.



The resulting alpha-values where: $0, 1, 1.5, 3.5$ and 4 (but not in this order). Decide which alpha-value is associated with each of the data points:

a) Positive sample $(-1, 1)$

b) Positive sample $(1, 1)$

c) Negative sample $(0, -1)$

d) Negative sample $(0, -2)$

e) Negative sample $(2, 1)$

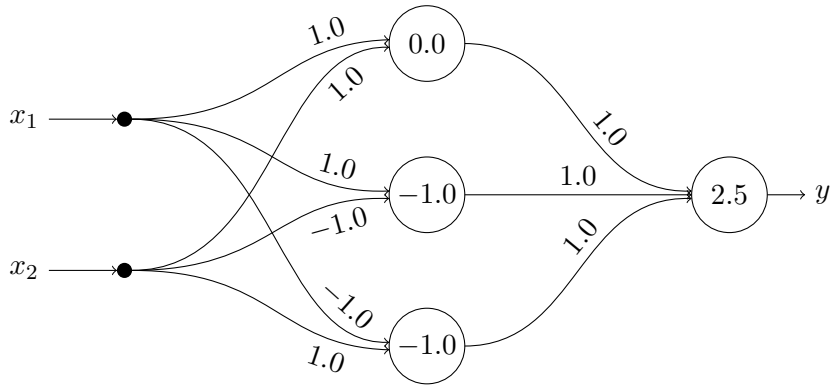
For each answer, you have to give a clear motivation, but you do not have to mathematically derive the answers.

Solution:

- Point **d** is outside the margin, so it must have $\alpha = 0$
- Point **b** is inside the margin, so slack is used and, hence, $\alpha = C = 4$
- Point **a** must have $\alpha = 1$ because the other possible values (1.5 and 3.5) would not make it possible to fulfill the constraint $\sum t_i \alpha_i = 0$ (the alphas for the two positive samples would then sum up to more than the sum of the alpha values left for the negative samples).
- The remaining question is which of points **c** and **e** should have $\alpha = 1.5$ and which should have $\alpha = 3.5$. We can note that **e** is closer to the decision boundary, but more importantly, it must balance point **b** on the other side, and that point has a maximally high alpha. Hence, **e** must be the bigger of the two: Point **e** has $\alpha = 3.5$.
- Point **c** gets the only remaining value: $\alpha = 1.5$.

B-8 Artificial Neural Networks (3p)

Consider a feed-forward neural network with threshold units. The number of nodes and all weight values are shown in the figure. Circles indicate threshold units with the threshold value written inside and output $\in \{-1, 1\}$. The small filled circles are just "pass through" nodes.

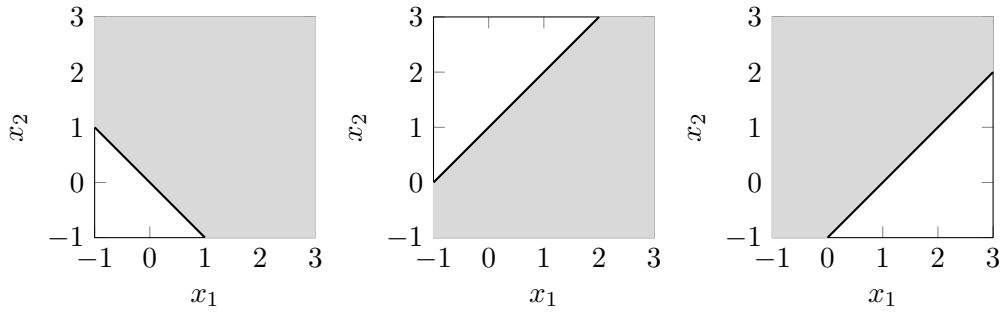


a) Draw a diagram of the input space and show the position of the separating hyperplanes implemented by the *hidden units*.

b) In the same figure, indicate the area where the output will be high ($y = 1$) by shading it.

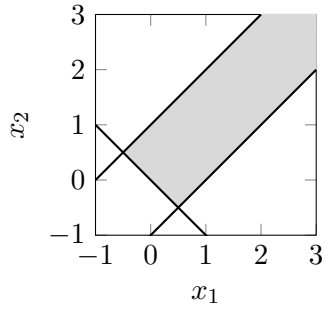
Solution:

a) The weights of the inputs to each of the hidden units defines the location of its corresponding separating hyperplane. In the 2D input space, this will correspond to these straight lines:



Here we have indicated with gray shading on which side the output is high.

b) The output unit will be above threshold only when all three hidden units are above threshold. This gives us this combined area:



B-9 Dimentionality Reduction (2p)

Given a set of feature vectors which all belong to a specific class C (i.e. with an identical class label), we performed PCA on them and generated an orthonormal basis $\{\mathbf{u}_1, ..., \mathbf{u}_p\}$ as the outcome. When the training samples in the class are well localised, the basis can be considered as a tool to represent possible variations of feature vectors within C in terms of a p-dimensional subsapce, $\mathcal{L}$. Provide an answer to the following questions.

a) We have a new input vector $\mathbf{x}$ whose class is unknown, and consider its projectiton length on $\mathcal{L}$. Represent the projection length by a simple formula.

b) Now, we consider a K-class classification problem and assume that a subspace $\mathcal{L}^{(j)}$ ($j = 1, ..., K$) has been computed with training data for each class, respectively. Briefly explain the way to determine the class to which vector $\mathbf{x}$ should belong using the projection lengths.

Solution:

a) $\sqrt{S}$ where $S = \Sigma_{i=1}^{p}(\mathbf{x}, \mathbf{u}_i)^2$

b) $\mathbf{x}$ should belong to the class where the projection length to the corresponding subspace is maximised.