KTH Computer Science and Communication

Exam in DD2421 Machine Learning
2019-03-08, kl 8.00 – 12.00

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 are the model parameters in a multivariate normal (Gaussian) distribution?

a) Mean vector and covariance matrix.

b) Number of data points.

c) Likelihood function.

Solution: a

A-2 Naive Bayes Classifier

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

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

b) All features are regarded as conditionally independent.

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

Solution: b

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}:

a) The entropy does not explain the uncertainty.

b) The entropy is smaller than one bit.

c) The entropy is equal to one bit.

Solution: b

A-4 Regularization

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) k-NN regression.

b) Ridge regression.

c) Logistic regression.

Solution: b

A-5 Artificial Neural Networks

What is the advantage of using a *multi-layered* artificial neural network (as opposed to a single-layered)?

a) Learning is guaranteed to converge to a unique solution

b) All input variables become independent

c) More complex decision bounderies can be formed

Solution: c

A-6 Support Vector Machine

What are the *support vectors* in a support vector machine?

a) Weights describing how important each sample is

b) Training data samples used to define the decision boundary

c) Orthogonal base vectors used to describe the kernel

Solution: b

A-7 Ensemble Learning

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

a) Ensemble methods are aimed to exploit a large number of training data.

b) Diverse models are trained and combined.

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

Solution: b

A-8 Principal Component Analysis (PCA)

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

a) To treat the data in infinite dimensional space.

b) To reduce the effective number of variables.

c) To find the least squares fit.

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) Bagging

b) Posterior probability

c) Dropout

d) Expectation Maximization

e) Curse of dimensionality

f) Occams's razor

g) k-means

h) RANSAC

1) A principle to choose the simplest explanation

2) Probability before observation

3) Algorithm to learn with latent variables

4) Issues in data sparsity in space

5) Estimating expected value

6) An approach to train artificial neural networks

7) Random strategy for amplitude compensation

8) A strategey to generate k different models

9) Probability at a later time

10) Method for estimating the mean of k observations

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

12) A technique for assessing a model while exploiting available data for training and testing

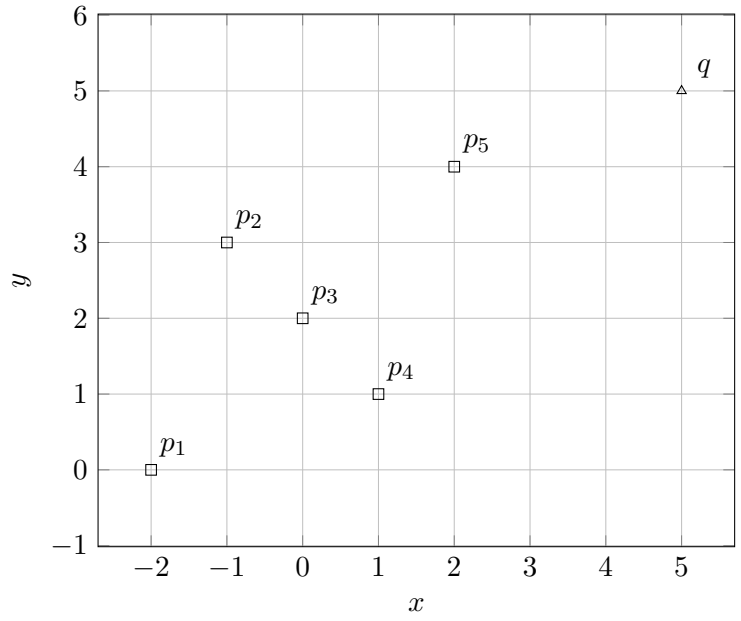
13) Conditional probability taking into account the evidence

14) Clustering method based on centroids

15) Bootstrap aggregating

Solution: a-15, b-13, c-6, d-3 e-4, f-1, g-14, h-11

B-2 Probabilistic regression (3p)



We want to perform probabilistic linear regression assuming Gaussian noise with zero mean and variance σ^2. The free parameters in the model are fit according to the Maximum Likelihood criterion to the data points $\{p_1, \ldots, p_5\}$ indicated by square markers in the figure.

a) Give the mathematical definition of the model and indicate which are the free parameters that need to be fit to the data.

b) What is the optimal value for the parameters?

c) Given the optimal parameters from the previous point, what is the likelihood ratio $\frac{\mathcal{L}(q)}{\mathcal{L}(p_3)}$ between point q indicated by the triangle marker in the figure and p_3?

Motivate your answers! Numerical answers without motivation give zero points.

Solution:

a) The full model is defined by:

$$y = w_0 + w_1 x + \epsilon,$$

with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The free parameters that we optimize are just w_0 and w_1 because σ^2 is assumed to be given. The corresponding posterior for y given x is a Gaussian distribution with parameters:

$$y|x \sim \mathcal{N}(w_0 + w_1 x, \sigma^2).$$

b) Maximizing the likelihood in the presence of Gaussian error with zero mean corresponds to minimizing the sum of square errors:

$$\text{SSE} = \sum_{i=1}^{5} e_i^2 = \sum_{i=1}^{5} (w_0 + w_1 x_i - y_i)^2.$$

The optimal model parameters are obtained by differentiating the above expression with respect to w_0 and w_1 and setting the derivatives to zero.

Plugging in the values of our data points we obtain:

$$\begin{aligned}
\text{SSE} &= (w_0 + w_1(-2) - 0)^2 + && (p_1 = (-2, 0)) \\
&\quad (w_0 + w_1(-1) - 3)^2 + && (p_2 = (-1, 3)) \\
&\quad (w_0 + w_1(0) - 2)^2 + && (p_3 = (0, 2)) \\
&\quad (w_0 + w_1(1) - 1)^2 + && (p_4 = (1, 1)) \\
&\quad (w_0 + w_1(2) - 4)^2 = && (p_5 = (2, 4)) \\
&= (w_0 - 2w_1)^2 + \\
&\quad (w_0 - w_1 - 3)^2 + \\
&\quad (w_0 - 2)^2 + \\
&\quad (w_0 + w_1 - 1)^2 + \\
&\quad (w_0 + 2w_1 - 4)^2.
\end{aligned}$$

We need to differentiate the above expression with respect to w_0 and w_1. Instead of simplifying the expression, because the derivative is a linear operator, we can differentiate each of the above terms and sum the derivatives.

Starting with w_0:

$$\begin{aligned}
\frac{\partial \text{SSE}}{\partial w_0} &= w_0 - 2w_1 + \\
&\quad w_0 - w_1 - 3 + \\
&\quad w_0 - 2 + \\
&\quad w_0 + w_1 - 1 + \\
&\quad w_0 + 2w_1 - 4 = \\
&= 5w_0 - 10.
\end{aligned}$$

Setting this derivative to zero we obtain $w_0 = 2$. Now differentiating with respect to w_1 we obtain:

$$\begin{aligned}
\frac{\partial \text{SSE}}{\partial w_1} &= -2(w_0 - 2w_1) + \\
&\quad -(w_0 - w_1 - 3) + \\
&\quad 0 + \\
&\quad w_0 + w_1 - 1 + \\
&\quad 2(w_0 + 2w_1 - 4) = \\
&= 10w_1 - 6.
\end{aligned}$$

Again, setting the derivative to zero we obtain $w_1 = \frac{3}{5}$.

c) In order to compute the likelihood ratio we should plug in the x and y values of the points q and p_3 into the formula for the probability density function of y given x. However, looking at the parameters for our linear model, we notice that the line

$$y = 2 + \frac{3}{5}x$$

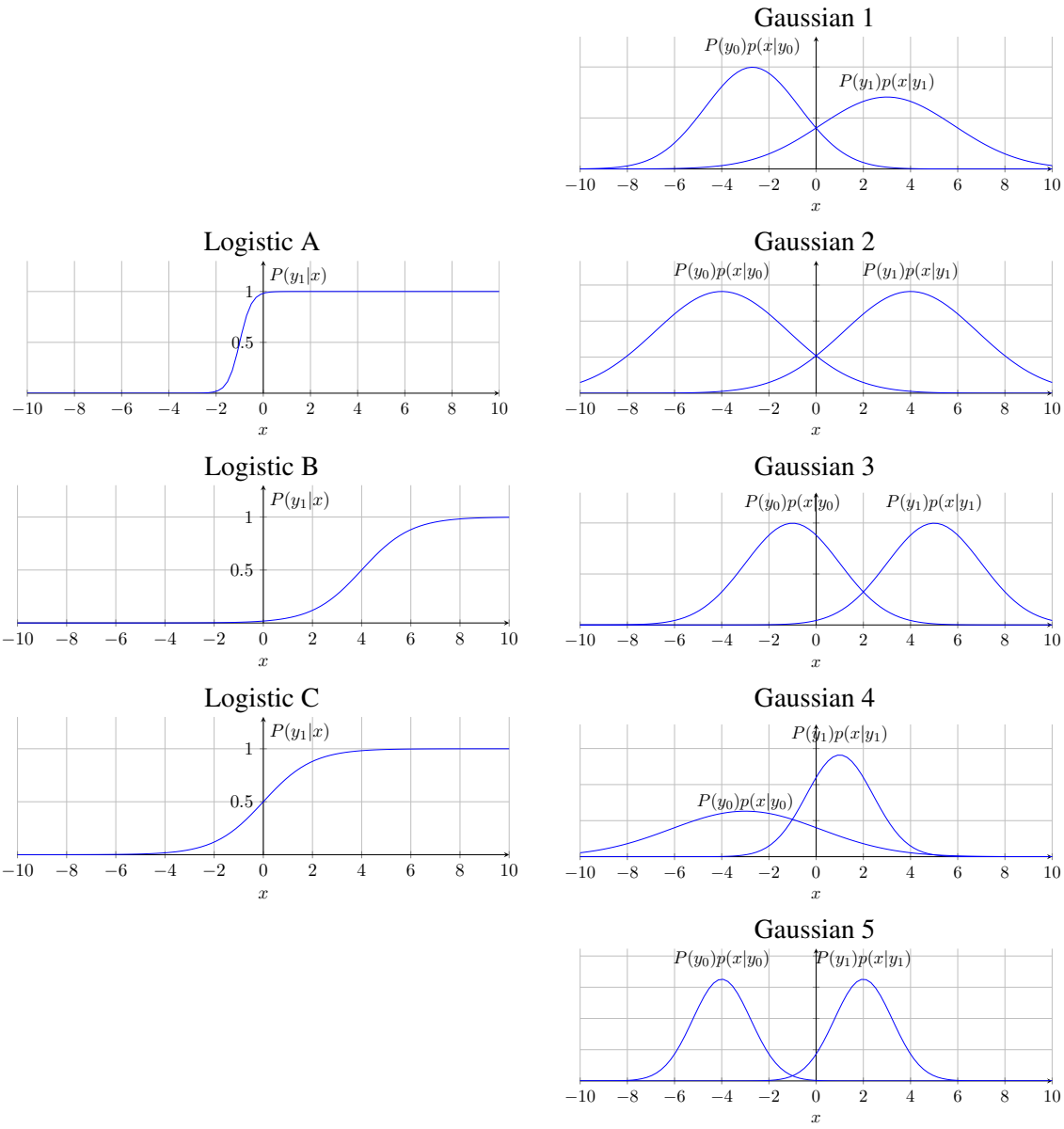
passes through both points. Because the likelihood is only dependent on the vertical distance of the point to the line, that is $e_i = 2 + \frac{3}{5}x_i - y_i$ which is zero in both cases, we know that the likelihood must be equal for the two points and the likelihood ratio must be 1.

B-3 Probabilistic classification (3p)

The figures below show eight binary classifiers. The three classifiers in the left column are based on logistic regression, and the plots show the posterior $P(y_1|x)$ for class y_1. The five classifiers in the right column are based on Gaussian distributions. In this case the plots show the *scaled posteriors* $P(y_i)p(x|y_i) = P(y_i)\mathcal{N}(\mu_i, \sigma_i^2)$ for each class y_i with $i = \{0, 1\}$.

For each logistic regression classifier, your task is to determine which Gaussian classifier, if any, may define the same posterior for class y_1. In other words, you need to assign each logistic regression classifier either to none or to one Gaussian classifier.

Motivate all your answers! Answers without motivation give zero points.



Solution: The posterior for class y_1 for the logistic regression is defined as:

$$P(y_1|x) \quad = \quad \text{sig}(w_0 + w_1 x) = \frac{1}{1 + \exp(-w_0 - w_1 x)},$$

where w_0 and w_1 are the model parameters and sig() is the sigmoid function. For the Gaussian classifiers, the posterior can be written, using the scaled posteriors, as:

$$P(y_1|x) = \frac{P(y_1)p(x|y_1)}{P(y_0)p(x|y_0) + P(y_1)p(x|y_1)} = \frac{1}{1 + \frac{P(y_0)p(x|y_0)}{P(y_1)p(x|y_1)}} = \frac{1}{1 + \frac{P(y_0)\mathcal{N}(\mu_0,\sigma_0^2)}{P(y_1)\mathcal{N}(\mu_1,\sigma_1^2)}}$$

Because of the formulation of the Gaussian distribution, it can be shown that this posterior is equivalent to the above sigmoid function when the prior probabilities for the two classes are equal ($P(y_0) = P(y_1)$), and the variance for the two class conditional distributions is the same ($\sigma_0^2 = \sigma_1^2 = \sigma^2$). Gaussian 1 and 4 are therefore excluded because they clearly do not satisfy these conditions.

We can also notice that the posterior is equal to 0.5 when the two scaled posteriors are equal, that is, where the two Gaussian distributions meet. For this reason:

- Logistic A is compatible with Gaussian 5 (meeting at $x = -1$),
- Logistic C is compatible with Gaussian 2 (meeting at $x = 0$),
- Logistic B does not have any corresponding Gaussian classifier (no pair of Gaussian distributions meet at $x = 4$).

In order to be sure about the equivalence in the first two cases, we would have to compare the exact model parameters (w_0 and w_1 for logistic regression and μ_0, μ_1 and σ^2 for the Gaussian classifiers), and to make sure that the following equalities are satisfied:

$$w_0 = \frac{\mu_0^2 - \mu_1^2}{2\sigma^2},$$
$$w_1 = \frac{\mu_1 - \mu_0}{\sigma^2}.$$

However, this is beyond the scope of the question and it is harder to verify from the plots.

B-4 Classification (3p)

Suppose that we take a data set, divide it into two parts of equal size, Part I and Part II. We try out two different classification procedures, by using Part I and Part II as our training set and test set, respectively, That is, we use half of the data for training, and the remaining half for testing. First we use 1-Nearest Neighbor rule (1-NN) and get an average error rate (averaged over both test and training data sets) of 8%. Next we use Logistic Regression and get an error rate of 12% on the training data. We also get the average error rate (averaged over both test and training data sets) of 13%.

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

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

c) Now, we swap the roles of Part I and Part II, and repeat the same experiments. On the test set (Part I), we get an error rate of 12% with both 1-NN and Logistic Regression. Based on all these results, by the cross-validation, indicate the method which we should prefer to use for classification of new observations, with a simple reasoning.

Solution:

a) 16%. (Training error for 1-NN is always zero, and therefore the testing error is 16%.)

b) 14%.

c) Logistic Regression because it achieves lower error rate on the test data on average (13% < 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**) and **b**), indicate which of {i,ii,iii,iv,v} is correct, and *briefly justify your answers*.

a) As we *increase* s from 0, the *test* 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 variance and (squared) bias, respectively.

c) The LASSO is known to yield *sparse models*. Explain what property of it enables the sparcity in a short sentence.

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) v.

b) ii for variance, iii for bias.

The variable selection property.

B-6 Ensemble Methods (2p)

Briefly answer the following questions regarding ensemble methods of classification.

a) What are the two kinds of randomness involved in the design of Decision 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, and how are they 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. The weight is increased if misclassified, and decreased if classified correctly.
The reliability is then used as the coefficient.

B-7 Dimensionality reduction (3p)

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 $\mathcal{L}$. Describe how the projection length is represented, using 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.

Next, given a training set of feature vectors with binary labels, we consider to find a good dimension (direction) for their binary classification by using the concepts of *between-class variance*, σ_B^2, and averaged *within-class variance*, σ_W^2, of the feature vectors viewed along that direction.

a) What factor should be maximised in the most effective dimension for separating the two classes? Provide an answer by a simple formula including σ_B^2 and σ_W^2.

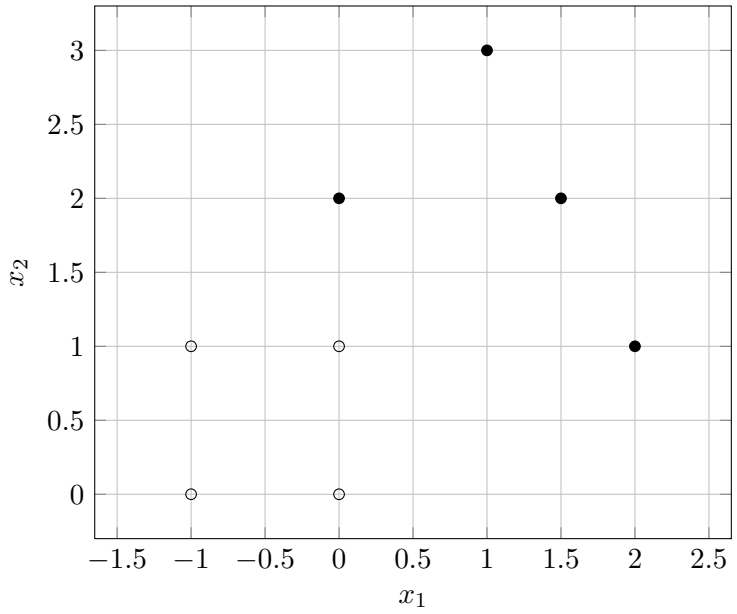
Solution:

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.

c) σ_B^2/σ_W^2.

B-8 Support Vector Machines (3p)



The eight datapoints in the figure are used to train a support vector machine. Filled circles are positive samples and unfilled circles are negative samples.

When a quadratic kernel ($\mathcal{K}(\vec{x}, \vec{y}) = (\vec{x} \cdot \vec{y} + 1)^2$) was used, and no slack was allowed ($C = \infty$), this resulted in three support vectors, with the corresponding α-values $\frac{1}{4}$, $\frac{1}{16}$, and $\frac{3}{16}$.

To classify new points, the normal indicator function is used:

$$\text{ind}(\vec{x}) = \sum_k t_k \alpha_k \mathcal{K}(\vec{x}, \vec{s}_k) - b$$

a) What will the three support vectors be? You do not need to calculate this mathematically; use you knowledge about support vectors to deduce which three points must be support vectors.

b) Which α-value must be associated with each of the three support vectors? (Give short motivations)

c) What value for b should be used in the indicator function?

Solution:

a)

$$s_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad s_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix} \qquad s_3 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

The decision boundary has to pass between the two classes, presumably bending downwards towards the right side. Removing any one of these three datapoints would clearly change the natural position of the decision boundary. Therefore, they must all be support vectors, since these are the points that define the boundary.

b)

$$\alpha_1 = \frac{1}{4} \qquad \alpha_2 = \frac{3}{16} \qquad \alpha_3 = \frac{1}{16}$$

$\sum_k t_k \alpha_k = 0$ means that the single negative support vector (s_1) must have the highest α-value. Of the other two, the point closest to the opposite class (s_2) will have the larger α.

c) The indicator function value is known to be 1 or -1 on the margin border, where the support vectors are (since we have no slack).

Using $\text{ind}(s_1) = -1$ (this is a negative point) we can calculate b as

$$b = \sum_k t_k \alpha_k \mathcal{K}(s_1, s_k) + 1 = -\frac{1}{4} \cdot 4 + \frac{3}{16} \cdot 9 + \frac{1}{16} \cdot 4 + 1 = \frac{31}{16} \approx 1.9$$

Using s_2 will instead give $b = 2$ which is equally correct as an answer here. The difference is due to rounding errors (the given α-values are not exact).

B-9 BackProp Learning (3p)

There are a number of values involved when training a layered feed-forward artificial neural network:

1) The number of hidden layers

2) Initial weights

3) Updated weights

4) Input vectors

5) Target values

6) Output values for the nodes

7) Local (generalized) errors

8) The number of nodes in each layer

9) The step-size

10) The number of training samples

Learning using the *back-propagation* (BackProp) learning algorithm involves computation in three steps; a forward propagating step, a backward propagating step, and a final local step. What values (pick one from the list above) is computed in each of these steps:

a) The forward propagating step

b) The backward propagating step

c) The local step

Solution:

a) The forward propagating step

Output values for the nodes.

This must be executed in a forward propagating manner, because the output of every node depends on the outputs from nodes in the preceding layer.

b) The backward propagating step

Local (generalized) errors.

The generalized errors are first computed at the output nodes, where the targets are known, and then propagated backwards to assign local errors to every node in the network.

c) The local step

Updated weights.

Once the output values (from **a**) and generalized errors (from **b**) are known everywhere in the network, the weights can be updated. This can be done locally since no more global communication is needed.