KTH Computer Science and Communication

Exam in DD2421 Machine Learning
2018-10-22, 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 give the correct answer to almost *all* questions. Only *one* error will be accepted, so pay good attention here.

A-1 Decision Trees

What principle is commonly used when building a decision tree?

a) Choose features that maximize information gain.

b) Maximize the tree height.

c) Minimize the number of leaf nodes.

Solution: a

A-2 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-3 Probabilistic Learning

What are latent variables in estimation?

a) Variables that do not influence the accuracy.

b) Variables that are not directly observed.

c) Deterministic factors.

Solution: b

A-4 Naive Bayes Classifier

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

a) All features are regarded as conditionally independent.

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

c) Prior probabilities are available.

Solution: a

A-5 Artificial Neural Networks

What is the underlying principle when using *backpropagation* to train an artificial neural network?

a) The weights are modified to minimize the mismatch between the actual and the desired output.

b) A Gaussian distribution is used to approximate the training data.

c) The number of hyper-planes is maximized using a dual formulation.

Solution: a

A-6 Support Vector Machine

What does it mean when a training sample gets a high α-value in a support vector machine?

a) That sample has a large influence on the resulting classifier.

b) That sample is associated with high uncertainty.

c) That sample occurs many times in the training set.

Solution: a

A-7 Ensemble Learning

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

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

b) The models to be combined should be as similar as possible to each other.

c) Weak models are trained and combined.

Solution: c

A-8 The Subspace Method

For the subspace methods, a technique of dimentionality reduction is often used to represent the data distribution in each class. Which of these techniques is most suited for this purpose?

a) Pulse-Code Modulation (PCM).

b) Phase Change Memory (PCM).

c) Principal Component Analysis (PCA).

Solution: c

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

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) Posterior probability

b) k-fold cross validation

c) Curse of dimensionality

d) The Lasso

e) Dropout

f) Perceptron Learning

g) RANSAC

h) Projection length

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

2) Method to find separating hyperplanes

3) A strategey to generate k different models

4) The last solution

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

6) Learning trying to mimic human vision

7) Conditional probability taking into account the evidence

8) A similarity measure in the subspace method

9) The length of cast shadow

10) Issues in data sparsity in space

11) Probability at a later time

12) An approach to train artificial neural networks

13) Sudden drop of performance

14) Random strategy for amplitude compensation

15) An approach to regression that results in feature seletion

Solution: a-7, b-5, c-10, d-15, e-12, f-2, g-1, h-8

B-2 Probabilistic classification (3p)

Consider a two-class classification problem in two dimensions $\mathbf{x} = (x_1, x_2)$. In the next page, you see four decision boundaries (A-D), and five maximum a posteriori classifiers based on Gaussian distributions with mean μ_i and covariance matrix Σ_i for each class i:

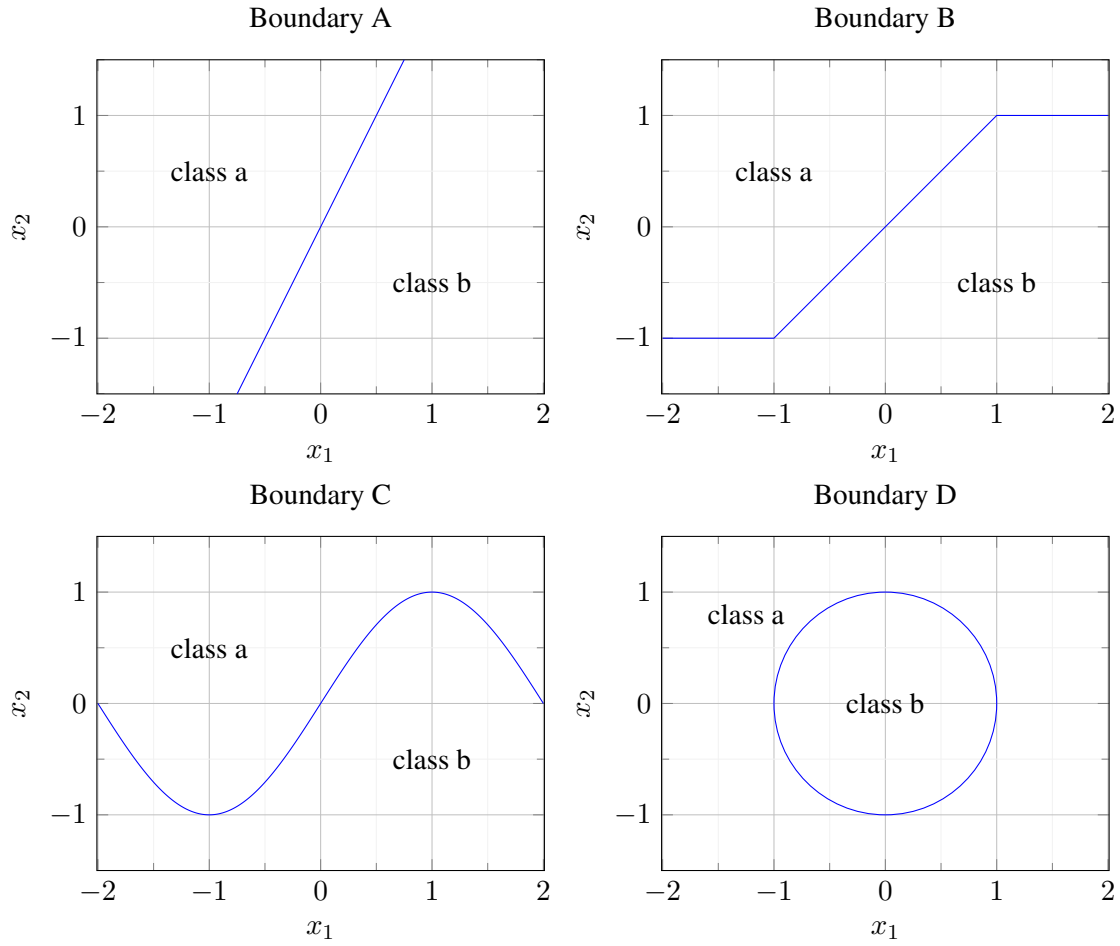
$$p(\mathbf{x}|\text{class } i) = \mathcal{N}(\mathbf{x}, \mu_i, \Sigma_i) = \frac{1}{2\pi\sqrt{\det \Sigma_i}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i)\right).$$

Consider equal a priori probabilities for each class, that is: $P(\text{class a}) = P(\text{class b}) = 0.5$, in all cases.

Your task is to assign, when appropriate, each decision boundary to the corresponding classifier(s), and motivate your choice. No complex derivations are required in your motivations, but you should be rigorous in your arguments.

Notes:

1. a correct assignment without motivation gives zero points,
2. incorrect assignments give negative points,
3. explaining why a certain decision boundary does not correspond to any model gives positive points,
4. the total number of points is at most 3 and cannot be negative (if you give more incorrect than correct answers, you will receive zero points in total).



Classifier 1

$$\mu_a = \begin{bmatrix} -1 \\ 0.5 \end{bmatrix} \quad \Sigma_a = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \mu_b = \begin{bmatrix} 1 \\ -0.5 \end{bmatrix} \quad \Sigma_b = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Classifier 2

$$\mu_a = \begin{bmatrix} -2 \\ 1 \end{bmatrix} \quad \Sigma_a = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix} \quad \mu_b = \begin{bmatrix} 2 \\ -1 \end{bmatrix} \quad \Sigma_b = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$$

Classifier 3

$$\mu_a = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \Sigma_a = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \mu_b = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \quad \Sigma_b = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Classifier 4

$$\mu_a = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma_a = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \mu_b = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma_b = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}$$

Classifier 5

$$\mu_a = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \Sigma_a = \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix} \quad \mu_b = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \quad \Sigma_b = \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}$$

Solution: In case of equal priors, the maximum a posteriori classifier is equivalent to the maximum likelihood classifier, that is, for a certain $\mathbf{x}'$ we choose class a over class b if $p(\mathbf{x}'|\text{class a}) > p(\mathbf{x}'|\text{class b})$. Because the Gaussian distributions are in the exponential family, it is convenient to do the comparison in log domain. In this domain, after some passages, we see that we need to check if the exponent of the first distribution minus the exponent of the second distribution is greater, equal or smaller than a constant. The decision boundary consists of all points for which the equality is true. Because both exponents are quadratic with respect to $\mathbf{x}$, the decision boundary can at most be a conic curve, that is: straight line (this is a degenerate case), circle, ellipsis, parabola or hyperbole. This is true in general, but requires some linear algebra to figure out from the matrix products in the exponents. In all the classifiers given in this problem, however, this property is easier to verify because all the covariance matrices are diagonal, which means that the exponents for class i simplify to terms in the form:

$$-\frac{(x_1-\mu_{i1})^2}{2\sigma_{i1}^2}-\frac{(x_2-\mu_{i2})^2}{2\sigma_{i2}^2},$$

and the decision boundary are be given by the expression

$$\overbrace{-\frac{(x_1-\mu_{a1})^2}{\sigma_{a1}^2}-\frac{(x_2-\mu_{a2})^2}{\sigma_{a2}^2}}^{\text{class a}}+\overbrace{\frac{(x_1-\mu_{b1})^2}{\sigma_{b1}^2}+\frac{(x_2-\mu_{b2})^2}{\sigma_{b2}^2}}^{\text{class b}}=\text{constant}. \tag{1}$$

If you expand the expression you will find only terms in the form x_1^2, x_1, x_2^2, x_2, and constants, which again correspond to conic curves. Following this argument, we can exclude both Boundary B, which is piece-wise linear, and Boundary C, which is sinusoidal.

For Boundary A, which is linear, we need to verify when the above expression only contains linear terms: x_1, x_2 and constants, or, similarly, when all the quadratic terms cancel out. If we only write the quadratic terms, we obtain:

$$-\frac{x_1^2}{\sigma_{a1}^2}-\frac{x_2^2}{\sigma_{a2}^2}+\frac{x_1^2}{\sigma_{b1}^2}+\frac{x_2^2}{\sigma_{b2}^2}=0,$$

which is clearly verified for each (x_1, x_2) when $\sigma_{a1}=\sigma_{b1}$ and $\sigma_{a2}=\sigma_{b2}$. Although not required to answer the question, this property is also valid for full covariance matrices: if $\Sigma_a=\Sigma_b$ the decision boundary is linear. It is left to you to prove this, if you like. Also, it can be verified that the constant on the right side of Eq.1 is zero if $\Sigma_a=\Sigma_b$ because the multiplicative term in front of the exponential is the same for both distributions. This condition is verified by Classifier 1, 2 and 3. In all three cases, the distributions also have the same spread in x_1 and x_2, which means that the expression above further simplifies and we can say that the point on the boundary are those equidistant from the means, that is, all $\mathbf{x}=(x_1, x_2)$ which verify:

$$(x_1-\mu_{a1})^2+(x_2-\mu_{a2})^2=(x_1-\mu_{b1})^2+(x_2-\mu_{b2})^2.$$

Geometrically, it can be verified that Boundary A, having slope 2 is consistent with both Classifier 1 and 2, but not with Classifier 3 that has a horizontal decision boundary with $x_2=0$.

Finally, Boundary D can be obtained if the distribution for class b has a smaller spread than for class a. In this case, class a has a relatively flat distribution around the mean while class b will be much more peaky close to the mean. Without proving this mathematically, we can see that Classifier 4 is the only one that has this property.

Summary:

- Boundary A is linear as Classifier 1-3, but the slope is only consistent with Classifier 1 and 2,
- Boundary D may be obtained with Classifier 4,
- Boundary B and C cannot be obtained with Gaussian distributions.

B-3 Probabilistic linear regression (3p)

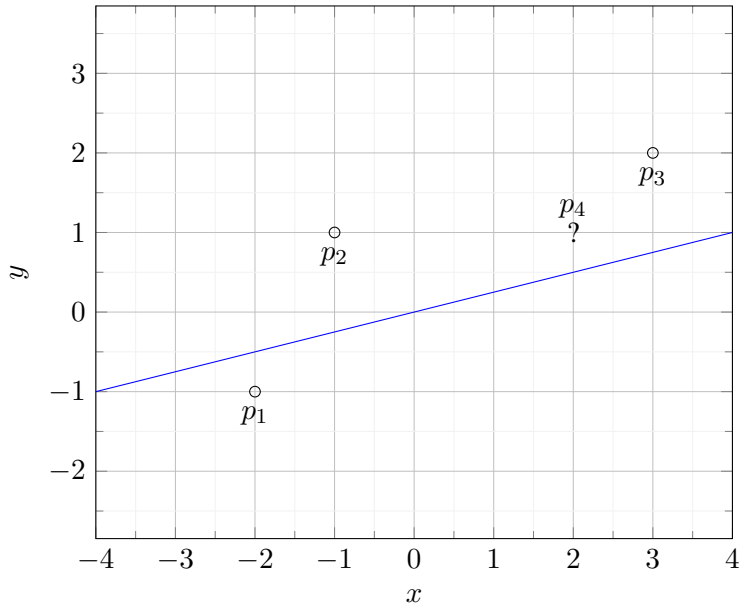
We consider a standard probabilistic linear regression model

$$y = ax + b + \epsilon,$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-0)^2}{2\sigma^2}\right),$$

where we force $b = 0$, that is, the line is forced to pass through the origin.

The figure below shows the regression model where the free parameter a has been fit through maximum likelihood estimation to four independent and identically distributed data points also shown in the figure. We call this specific value a_{ML}. We know both abscissa x_i and ordinate y_i only for three of the four points in the training data. These are shown in the figure as small circles. For the fourth point, only the abscissa $x_4 = 2$ is known, but we ignore the value of the ordinate y_4. To illustrate this, we use the symbol "?" for this point, on a location that does not necessary correspond to the true value of y_4.

Linear regression problem



a) inspecting the figure, report the value of the model parameters a_{ML},

b) calculate the ordinate y_4 of the missing point given the above assumptions and $a = a_{\text{ML}}$,

c) given the solution of the previous question, would the linear model in the figure still be optimal in the maximum likelihood sense if we also let b vary?

Solution:

a) we know that a is the slope of the line, so it is easy to see from the figure that $a_{\text{ML}} = \frac{1}{4}$.

b) Given the information from the question, the posterior of y given x and the model parameters is:

$$p(y|x, a, b, \sigma^2) = \mathcal{N}(ax, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-ax)^2}{2\sigma^2}\right),$$

that is, a Gaussian distribution with mean $\mu = ax$ and variance σ^2.

The maximum likelihood estimate is the value of a that maximizes the above quantity over the four points in our trainin data:

$$a_{\text{ML}} = \arg\max_a p(Y|X, a, b, \sigma^2) = \arg\max_a \prod_{i=1}^{4} p(y_i|x_i, a, b, \sigma^2),$$

where we have used the fact that the points are independent and identically distributed. Remember that the above expression is a posterior with respect of the regression problem, but it is a likelihood with respect of the model parameter (this is why it is a maximum likelihood estimate of the parameter a).

In logarithmic domain, after some simple steps, we find that maximizing the above expression is equivalent to minimizing the following:

$$a_{\text{ML}} = \arg\min_a \sum_{i=1}^{4} (y_i - ax_i)^2,$$

which corresponds to the fact that the maximum likelihood estimate is equivalent to the minimum sum of square error solution. To solve this optimization problem we set to zero the derivative of the above expression with respect to the parameter a[1]. Doing this we obtain:

$$\frac{\partial}{\partial a} \sum_{i=1}^{4} (y_i - ax_i)^2 = -2 \sum_{i=1}^{4} x_i(y_i - ax_i) = 0 \qquad \Longleftrightarrow \qquad \sum_{i=1}^{4} x_i y_i = a_{\text{ML}} \sum_{i=1}^{4} x_i^2.$$

From the figure we can read the abscissas for all points and the ordinates for three of the points. We have:

$$X = \{-2, -1, 3, 2\}$$
$$Y = \{-1, 1, 2, y_4\}.$$

We also know the value of $a_{\text{ML}} = \frac{1}{4}$ from the previous question. We can, therefore rewrite the right side of the expression as:

$$a_{\text{ML}} \sum_{i=1}^{4} x_i^2 = \frac{1}{4}\left((-2)^2 + (-1)^2 + 3^2 + 2^2\right) = \frac{18}{4} = \frac{9}{2},$$

[1] Note that this is different from setting the value of a to a_{ML} and optimize with respect to y_4: in this second case, the optimum point would always be on the line because we want to minimize the sum of square errors and the other errors are fixed.

and the left side of the expression, as:

$$\sum_{i=1}^{4} x_i y_i = -2 \cdot (-1) + (-1) \cdot 1 + 3 \cdot 2 + 2 \cdot y_4 = 7 + 2y_4.$$

Equating the two sides we obtain:

$$y_4 = -\frac{5}{4} = -1.25$$

c) if b is also able to vary, the optimization above becomes:

$$(a, b)_{\text{ML}} = \arg\min_{(a,b)} \sum_{i=1}^{4} (y_i - ax_i - b)^2.$$

If, by using $a = a_{\text{ML}} = \frac{1}{4}$, we optimize with respect to b and obtain $b_{\text{ML}} = 0$ we know that the model depicted in the figure is still optimal. Again, we can differentiate the expression with respect to b and obtain:

$$\frac{\partial}{\partial b} \sum_{i=1}^{4} (y_i - ax_i - b)^2 = -2 \sum_{i=1}^{4} (y_i - ax_i - b) = 0 \iff 4b = \sum_{i=1}^{4} y_i - a \sum_{i=1}^{4} x_i.$$

We can calculate $a_{\text{ML}} \sum_{i=1}^{4} x_i = \frac{1}{2} = 0.5$, but if we sum the ordinates for the four points we obtain $\sum_{i=1}^{4} y_i = \frac{3}{4} = 0.75$ and, therefore, the optimal value for b is not zero and the model in figure is not optimal any longer.

B-4 Classification (2p)

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 10%. Next we use the Subspace Method and get an error rate of 13% on the training data. We also get the average error rate (averaged over both test and training data sets) of 15%.

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 the Subspace Method on the test set? Based on these results, indicate the method which we should prefer to use for classification of new observations, with a simple reasoning.

Solution:

a) 20%. (Training error for 1-NN is always zero, and therefore the testing error is 20%.)

b) 17%. The Subpace Method because it achieves lower error rate on the test data (17% < 20%).

B-5 Ensemble Methods (3p)

Give answers to the following questions on Ensemble Learning.

a) Mainly two kinds of randomness are known to form the basic principle of Random Forests. In which two of the following processes are those randomnesses involved?

i. In combining the results from multiple trees.
ii. In generating bootstrap replicas.
iii. In the rule of terminating a node as a leaf node.
iv. In feature selection at each node.
v. In the way to compute the information gain.

Simply indicate two among those above.

b) In Adaboost algorithm, each training sample is given a weight and it is updated according to some factors through an iteration of training classifiers.

b-1. What are the two most dominant factors in updating the weights?
b-2. How are those two factors used to update the weight?

Solution:

a) ii and iv.

b) **b-1**) 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.
b-2) The weight is increased if the sample was misclassified, and decreased if correctly classified. The reliability is then used as the coefficient.

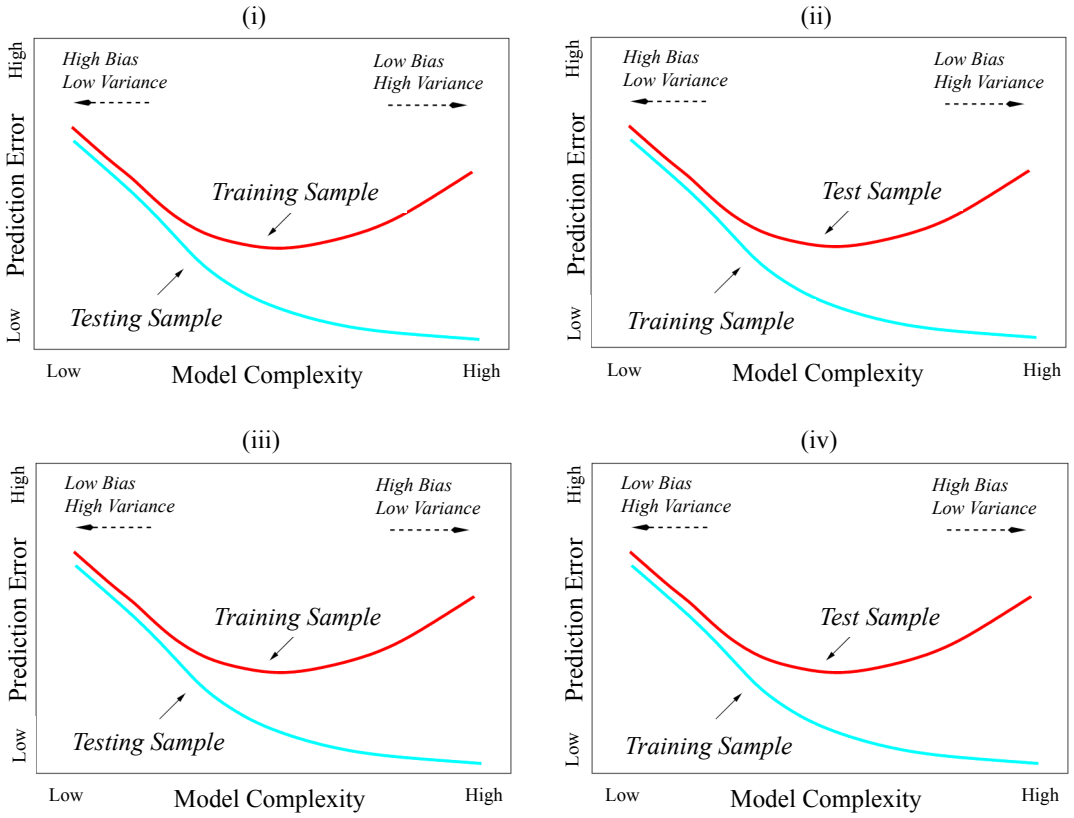


Figure 1. Typical behavior of prediction error plotted against model complexity.

B-6 Bias and Variance (3p)

a) One of the four subfigures (i)-(iv) in Figure 1 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) In ridge regression, relative to least squares, a term called *shrinkage penalty* is added in the quantity to be minimised. They give improved prediction accuracy in some situations. Briefly explain when this happens in terms of bias-variance trade-off.

Solution:

a) (ii)

b) Overfitting.

c) When the increase in bias is less than the decrease in variance.

B-7 Support Vector Machines (3p)

Consider a one-dimensional classification problem with a small training data set consisting of only three samples:

$$\vec{x}_1 = [1], \qquad \vec{x}_2 = [2], \qquad \vec{x}_3 = [3]$$

and the corresponding target classes:

$$t_1 = 1, \qquad t_2 = -1, \qquad t_3 = 1.$$

When a SVM is trained with these samples using a quadratic kernel, $\mathcal{K}(\vec{x}, \vec{y}) = (\vec{x}^T\vec{y} + 1)^2$, the resulting alpha values become

$$\alpha_1 = 11, \qquad \alpha_2 = 18, \qquad \alpha_3 = 7.$$

Classification of new data is done using the indicator function:

$$\text{ind}(\vec{s}) = \sum_i \alpha_i t_i \mathcal{K}(\vec{s}, \vec{x}_i) - b$$

a) When solving this, a matrix $P_{i,j} = t_i t_j \mathcal{K}(\vec{x}_i, \vec{x}_j)$ is used. What is the contents of this matrix for this particular example?

b) What is the value of b (used in the indicator function)?

c) What is the value of the indicator function for a new point at $\vec{x} = [0]$?

Solution:

a)

$$P = \begin{bmatrix} 1\cdot 1\cdot(1\cdot 1+1)^2 & 1\cdot -1\cdot(1\cdot 2+1)^2 & 1\cdot 1\cdot(1\cdot 3+1)^2 \\ -1\cdot 1\cdot(2\cdot 1+1)^2 & -1\cdot -1\cdot(2\cdot 2+1)^2 & -1\cdot 1\cdot(2\cdot 3+1)^2 \\ 1\cdot 1\cdot(3\cdot 1+1)^2 & 1\cdot -1\cdot(3\cdot 2+1)^2 & 1\cdot 1\cdot(3\cdot 3+1)^2 \end{bmatrix} = \begin{bmatrix} 4 & -9 & 16 \\ -9 & 25 & -49 \\ 16 & -49 & 100 \end{bmatrix}$$

b) We know that $\text{ind}(1) = 1$ since it is a support vector.

$$\text{ind}(1) = \sum_i \alpha_i t_i \mathcal{K}(1, \vec{x}_i) - b = 1 \quad \Rightarrow \quad b = 11\cdot 4 - 18\cdot 9 + 7\cdot 16 - 1 = -7$$

c)

$$\text{ind}(0) = 11\cdot 1 - 18\cdot 1 + 7\cdot 1 + 7 = 7$$

B-8 Artificial Neural Networks (3p)

For each of the learning algorithms **a**–**c**, what could be a possible cause for failure, that is, that the algorithm did not find a solution which classifies the training data correctly?

a) A single linear hyperplane classifier, trained using perceptron learning.

b) A two layer neural network with 10 hidden units, trained using backpropagation.

c) A support vector machine with a RBF kernel.

For each algorithm **a**–**c**, state which of the three alternative explanations below (**i**–**iii**) are possible causes of failure. Multiple answers may be correct (including none or all), so you must motivate why each alternative is possible or impossible (in total, 9 yes/no answers with short motivations).

i) The data was not linearly separable

ii) Learning got stuck in a local minimum

iii) Initial weights were inappropriate

Solution:

a) Perceptron learning

- **i**) Yes, a pure hyperplane can only do linear separation
- **ii**) No, will always converge
- **iii**) No, initial weights do not affect convergence

b) Backpropagation

- **i**) No, unless only a single hidden unit is used, linear separability is not necessary
- **ii**) Yes, BP is prone to converge to local minima
- **iii**) Yes, a bad starting position can ruin convergence

c) SVM

- **i**) No, the RBF kernel makes non-linear separation possible
- **ii**) No, the optimization problem is convex, so only one minima exists
- **iii**) No, there are no initial weights involved

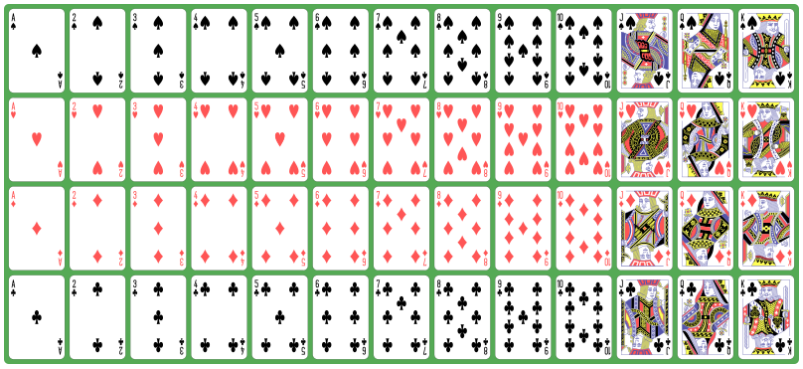


Figure 2. Playing cards consisting of 52 patterns.

B-9 Information Contents (3p)

Imagine that you are playing with Cards and randomly sample *three cards* out of the pile of 52 cards (see Figure 2) *with replacement*, i.e. you sequentially draw a card but return it to the pile each time you have seen what it is.

a) At each instance of drawing a card, what is the Shannon information content of the outcome with respect to the suit, one of {*Spades, Hearts, Diamonds, and Clubs*}, measured in bits?

b) You play a game with a rule that you win if the suits of all the *three* cards are of *the same colour*, either *black or red*. Otherwise you lose. How unpredictable is the outcome of this game (win or lose)? Answer in terms of entropy, measured in bits.

c) With respect to the outcome of the game in **b**), what is the expected information gain by drawing the first two card, i.e. by seeing (the suit colours of) the first card and the second card?

Note: if you do not have a calculator, answer with an expression but simplify it as much as possible.

Solution:
a) At each instance, it is $\log_2 \frac{1}{1/4} = 2$ (bits).
b) There are $2^3 (= 8)$ patterns in terms of the combinations of the colours (with equal probability). For two of these you win (all black or all red), for the remaining cases you lose.
Let $f(p_1, p_2) = -p_1 \log_2 p_1 - p_2 \log_2 p_2$ Entropy: $f(\frac{2}{8}, \frac{6}{8}) = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4} \approx 0.811$
c) Two senarios: the second card can bear the same colour as the first card, or different colour. These happen with probabilities, $\frac{1}{2}$ and $\frac{1}{2}$, respectively. If the first two bear different colours, we know we will lose and hence the remaining entropy is zero. If the first two are of the same colour, we still have half the chance for winnning/losing with the entropy being 1.
The information gain: $0.811 - (\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 1) = 0.311$