KTH Computer Science and Communication

Exam in DD2431 Machine Learning
2015-10-27, kl 14.00 – 18.00

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

A Questions for pass or fail

Only one alternative is correct for each question.

Note: To pass the exam you must give the correct answer on almost *all* questions in this section. Only *one* error will be accepted, so be very careful not to make any unnecessary mistakes here.

A-1 Probabilistic Learning

The log likelihood of a data set composed by independent and identically distributed observations can be computed

a) as the product of the log likelihoods of each individual observation.

b) as the sum of the log likelihoods of each individual observation.

c) as the integral of the probability distribution function.

d) using a conjugate prior.

Solution: b

A-2 Decision Trees

What principle is commonly used when building a decision tree?

a) Maximize the number of leaf nodes.

b) Maximize the projection length.

c) Maximize the distance to borderline samples.

d) Choose questions that maximize information gain.

Solution: d

A-3 Overfitting

You have trained a model (classifier) using some training sample data. Under which conditions is overfitting most likely to occur?

a) Few training samples are used to train a relatively simple model.

b) Many training samples are used to train a relatively simple model.

c) Few training samples are used to train a relatively complex model.

d) Many training samples are used to train a relatively complex model.

Solution: c

A-4 Kernel function

What is the role of the *kernel function* in a Support Vector Machine?

a) Perform scalar products in a virtual high-dimensional space.

b) Low-pass filter the input to reduce the effect of noise.

c) Increase the length of the weight vector.

d) Separate positive from negative samples.

Solution: a

A-5 Artificial Neural Networks

What is the main reason for using *hidden layers* in a feed-forward artificial neural network?

a) The network will generalize better.

b) More training data can be processed in parallel.

c) Non-linear decision boundaries can be learned.

d) Noisy data will be filtered away before reaching the classifier.

Solution: c

A-6 Regression and Classification

Which one of the following is true as the main difference between regression and classification?

a) Techniques of classification require sampling while those of regression require pruning.

b) The output is discrete for classification and continuous for regression.

c) Nearest neighbours can be used for regression but not for classification.

d) Classification is based on supervised learning whereas regression is based on unsupervised learning.

Solution: b

A-7 Ensemble Learning

Weak learners are trained and combined in different Ensemble methods in different manners. Which one of the following best describes the characteristics of the training process?

a) In Bagging and Decision Forests, the weak learners are trained in a parallel manner.

b) In Boosting and Bagging, the weak learners are trained in a parallel manner.

c) In Decision Forests and Boosting, the weak learners are trained in a parallel manner.

d) In Bagging, Boosting, and Decision Forests, the weak learners need be trained sequentially.

Solution: a

A-8 Principal Component Analysis (PCA)

Which one of the following statements is *incorrect*?

a) PCA serves for subspace methods to represent the data distribution in each class.

b) PCA is useful for reducing the effective dimensionality of data.

c) PCA can be seen as a tool for acquiring an approximated representation of data.

d) PCA is a supervised learning method that requires labeled data.

Solution: d

B Questions for higher grades

Preliminary number of points required for different grades:

$$\begin{aligned} 22 \leq p \leq 24 &\rightarrow A \\ 18 \leq p < 22 &\rightarrow B \\ 12 \leq p < 18 &\rightarrow C \\ 6 \leq p < 12 &\rightarrow D \\ 0 \leq p < 6 &\rightarrow E \end{aligned}$$

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) k-nearest neighbour

b) Support Vector

c) Perceptron learning

d) Categorical distribution

e) Subspace

f) Dropout

g) EM-algorithm

h) Variance

1) Location of concentration of a set of data

2) Samples on the margin of the decision surface

3) An approach to iteratively fitting model parameters with latent variables

4) A situation with short of samples in the training phase

5) A learning approach by ensemble machines

6) Measure of spread of a random variable

7) A subportion of area defined by two sets of parallel lines

8) Evidence for a specific hypothesis

9) A space spanned by a set of linearly independent vectors

10) Probability divided by the bias

11) Distribution of discrete stochastic variables

12) A method of regularization used in deep neural networks

13) Class prediction by a majority vote

14) Error driven method to compute weights in a single layer network

15) Learning based on distorted versions of the training data

Solution: a-13, b-2, c-14, d-11, e-9, f-12, g-3, h-6

B-2 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 Naïve Bayes classifier and get an error rate of 20% on the training data. We also get the average error rate (averaged over both test and training data sets) of 24%. Next we use 1-nearest neighbor and get an average error rate (averaged over both test and training data sets) of 16%.

a) What was the error rate with Naïve Bayes classifier 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.

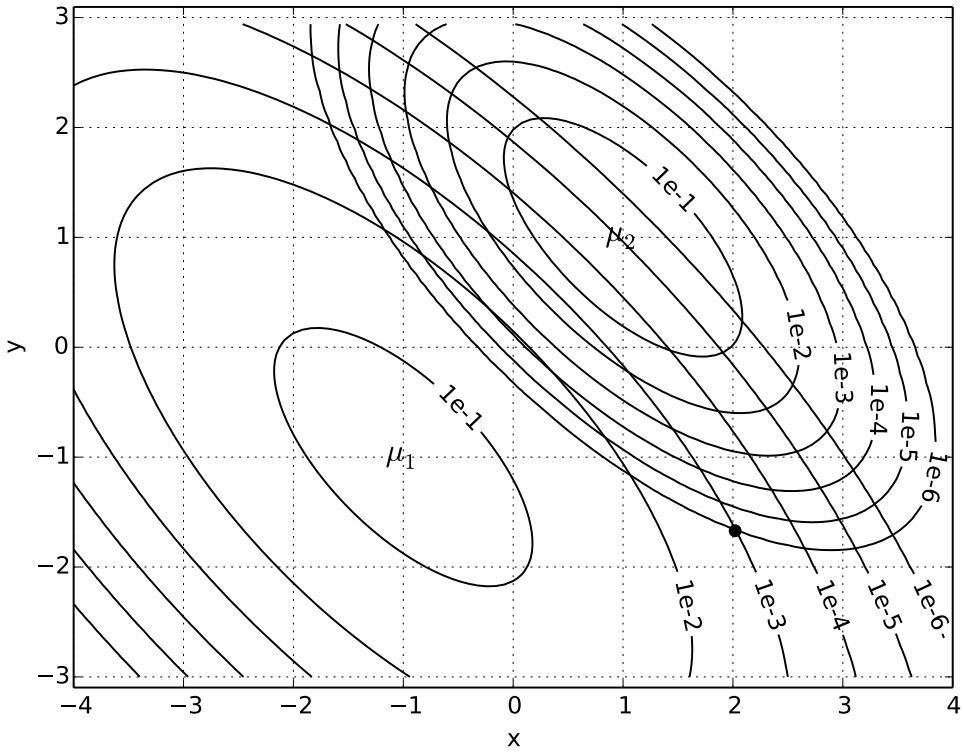


Figure 1. The two probability distribution functions (PDFs) from problem B-3. The mean values are indicated by μ_1 and μ_2. The contour lines correspond to the locations where the PDFs equal the following values: $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$.

Solution:
a) 28% b) Naïve Bayes classifier because it achieves lower error rate on the test data (28% < 32%).

B-3 Probability based learning (4p)

Figure 1 displays a two-class Maximum Likelihood (ML) classifier where the likelihood of a data point x given each class c is a two dimensional multivariate Gaussian (Normal) distribution,

$$P(x|c) = \mathcal{N}(x, \mu_c, \Sigma_c).$$

The mean vectors for class 1 and 2 are respectively

$$\mu_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \qquad \mu_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

a) looking at the figure, determine which of the two covariance matrices below corresponds to class 1 and 2 respectively, give a qualitative explanation for your answer:

$$\Sigma_a = \begin{bmatrix} 0.3 & -0.2 \\ -0.2 & 0.3 \end{bmatrix} \qquad \Sigma_b = \begin{bmatrix} 1.2 & -0.8 \\ -0.8 & 1.2 \end{bmatrix},$$

b) using Figure 1, classify the following points according to the ML classifier:

$$x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad x_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix} \qquad x_3 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \qquad x_4 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

c) is the separation boundary between the two classes linear (straight line)? Motivate your answer using Figure 1.

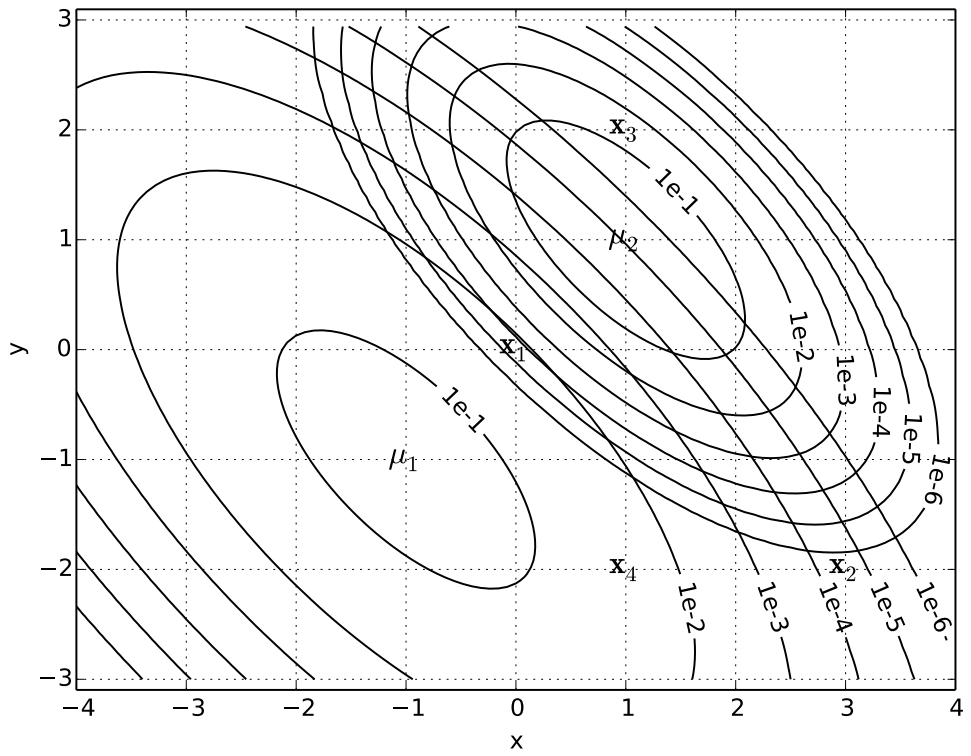
d) You want to classify the point indicated by the black circle in Figure 1 with a Maximum a Posteriori (MAP) classifier based on the likelihood functions above and the a priori probabilities of the classes $p_1 = P(c=1)$ and $p_2 = P(c=2)$. What is the range for p_1 and p_2 that will make the MAP classifier assign the point to class 2?

Note: no complex derivations are required to solve any of the above questions. If you need to refer to the figure to explain your solution, **do not draw on this sheet**: we will not receive it with your solutions.

Solution:

a) The two covariance matrices only differ by a scalar factor ($\Sigma_b = 4\Sigma_a$). Because the variance (diagonal) is a measure of spread, multiplying it by a factor > 1 means increasing the spread. You can therefore deduce from the figure that Σ_b corresponds to class 1 and Σ_a to class 2.[1]

b) In the figure below, we added the data points to Figure 1:

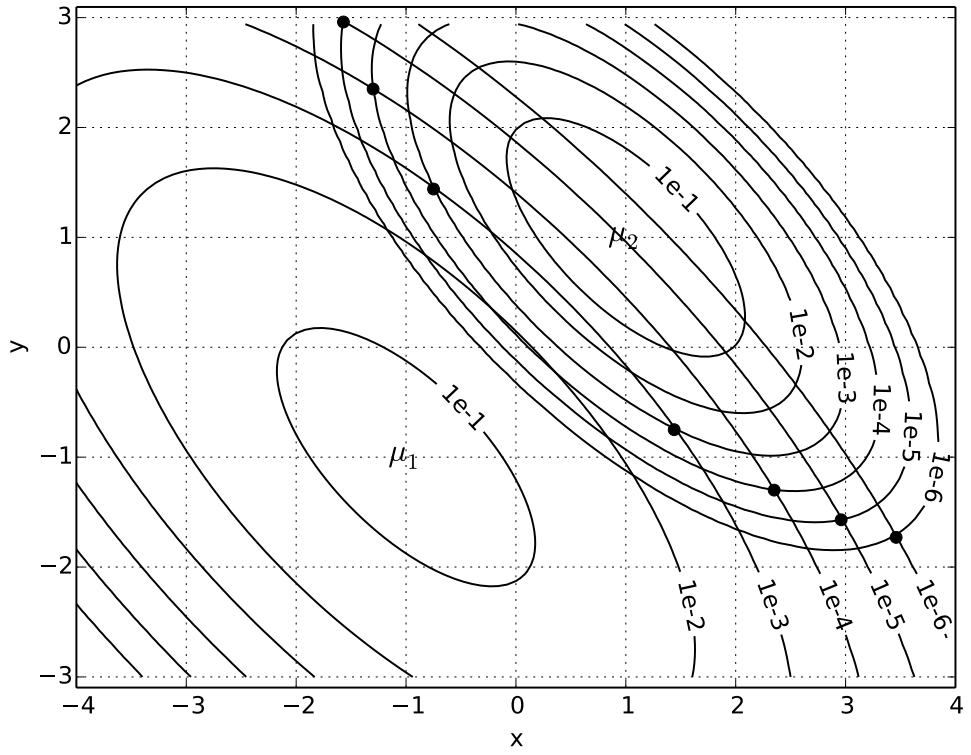


In order to classify each point we need to check whether $P(x_i|c=1) > P(x_i|c=2)$. From the figure we do not know the exact value of $P(x_i|c)$ in general, but we can deduce the range

[1] To be formally correct, the spread is related to the eigenvalues of the covariance matrix. However, if we multiply a matrix by a scalar, the corresponding eigenvalues are also multiplied by the same scalar.

of $P(x_i|c)$ from the contour lines for each class. For example x_1 is within the $\frac{1}{100}$ line for class 1, but it is outside the $\frac{1}{10000}$ line for class 2. This means that $P(x_1|c=1)$ is at least 100 times bigger than $P(x_1|c=2)$, and x_1 clearly belongs to class 1. With the same kind of arguments, we can say that x_2 and x_4 belong to class 1, whereas x_3 belongs to class 2.

c) To answer this question, you can find the intersections between the corresponding contour lines for the two distributions. All those points must sit on the separation line because $P(x|c=1) = P(x|c=2)$. We can see from the figure below that the separation line cannot be straight.



d) the point indicated in the figure sits on the $\frac{1}{1000}$ line for class 1 and on the $\frac{1}{1000000}$ line for class 2. This means that $P(x|c=1) = 1000\ P(x|c=2)$. In order to classify the point as class 2 using MAP we want the posterior for class 2 to be greater than the posterior for class 1, that is, $P(c=2|x) > P(c=1|x)$. This is verified if $P(x|c=2)P(c=2) > P(x|c=1)P(c=1)$, where we have disregarded the evidence $P(x)$ that is constant for both classes. We want, therefore

$$\frac{p_2}{p_1} = \frac{P(c=2)}{P(c=1)} > \frac{P(x|c=1)}{P(x|c=2)} = 1000.$$

That is:

$$\begin{cases} p_2 > 1000\, p_1, \\ p_1 + p_2 = 1, \\ 0 \leq p_1 \leq 1, \\ 0 \leq p_2 \leq 1. \end{cases}$$

That are verified by

$$\begin{cases} 0 \leq p_1 < \frac{1}{1001} \approx 0.001, \\ p_2 = 1 - p_1. \end{cases}$$

The interpretation of the above equations is that we want the prior information in favor of class 2 to at least as strong as the likelihood in favor of class 1.

B-4 Information Content (3p)

Imagine that you are playing with Cards and randomly sample three cards out of the pile of 52 cards *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 {*Clubs, Spades, Diamonds, Hearts*}, measured in bits?

b) You play a game with a rule that you win if all the *three* cards are of *different* suit from each other. 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 from seeing the suit of the first card?

d) Now, what is the expected information gain from also seeing the suit of 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 $4^3 (= 64)$ patterns in terms of the combinations of the suits (with equal probability). For 4! (=24) of these you win, 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: $f(\frac{24}{64}, \frac{40}{64}) = -\frac{3}{8} \log_2 \frac{3}{8} - \frac{5}{8} \log_2 \frac{5}{8} \approx 0.954$
c) Seeing the suit of the first card does not make any diffrence in the outcome. In fact, regardless of the suit of the first card, the entropy will be $f(\frac{6}{16}, \frac{10}{16})$ which is just the same value as in **b**); the information gain is therefore zero.
d) Two senarios: the second card can bear the same suit as the first card, or different suit. These happen with probabilities, $\frac{1}{4}$ and $\frac{3}{4}$, respectively.
If the first two bear the same suit, we know we will lose and hence the remaining entropy is zero.
If the first two are different, we still have half the chance of winnning, with the entropy being one.
The information gain: $0.954 - (\frac{1}{4} \cdot 0 + \frac{3}{4} \cdot 1) = 0.204$

B-5 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 taken into account?

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) the reliability of the weak classifier and (ii) if the sample was misclassified.

The weight is increased if misclassified, and decreased if classified correctly. The reliability is based on the training error; the smaller the training error, the greater the reliability.

B-6 Regression with regularization (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

$$\sum_{n=1}^{N} (y_n - \mathbf{w}^\top \mathbf{x}_n)^2 + \lambda \sum_{i=1}^{d} w_i^2$$

for a particular value of λ.

For parts **a**) through **d**), indicate which of i. through v. is correct. Briefly justify your answer.

a) As we increase λ from 0, the training error (resisual sum of squares, RSS) will:
i. Steadily increase.
ii. Steadily decrease.
iii. Remain constant.
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.

d) Repeat **a**) for (squared) bias.

Solution: **a**)-i, **b**)-v, **c**)-ii, **d**)-i

B-7 Support Vector Machines (3p)

Training a support vector machine using a quadratic kernel

$$\mathcal{K}(\vec{x}, \vec{y}) = (\vec{x}^T \vec{y} + 1)^2$$

has resulted in the following three support vectors:

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

The first is a positive sample with $\alpha = 0.25$ while the other two are negative samples where $\alpha = 0.1$.

a) Determine how the following *new* datapoints will be classified:

$$x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad x_2 = \begin{bmatrix} 0 \\ -0.1 \end{bmatrix} \qquad x_3 = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \qquad x_4 = \begin{bmatrix} 2.5 \\ 1 \end{bmatrix}$$

You must show the formulas and calculations used to arrive at your answer.

b) Draw a diagram of the input space and show the shape and position of the decision boundary.

Solution:

a) We use the indicator function to classify the new points:

$$\text{ind}(\vec{x}) = \sum_i \alpha_i t_i \mathcal{K}(\vec{x}, \vec{s_i}) = 0.25 \cdot \mathcal{K}(\vec{x}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}) - 0.1 \cdot \mathcal{K}(\vec{x}, \begin{bmatrix} -1 \\ -1 \end{bmatrix}) - 0.1 \cdot \mathcal{K}(\vec{x}, \begin{bmatrix} 1 \\ -1 \end{bmatrix})$$

Fill in the given values:

$$\text{ind}(x_1) = \text{ind}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}) = 0.25 \cdot 1 - 0.1 \cdot 1 - 0.1 \cdot 1 = 0.05$$

x_1 is classified as positive.

$$\text{ind}(x_2) = \text{ind}(\begin{bmatrix} 0 \\ -0.1 \end{bmatrix}) = 0.25 \cdot 0.81 - 0.1 \cdot 1.21 - 0.1 \cdot 1.21 = -0.0395$$

x_2 is classified as negative.

$$\text{ind}(x_3) = \text{ind}(\begin{bmatrix} 2 \\ 1 \end{bmatrix}) = 0.25 \cdot 4 - 0.1 \cdot 4 - 0.1 \cdot 4 = 0.2$$

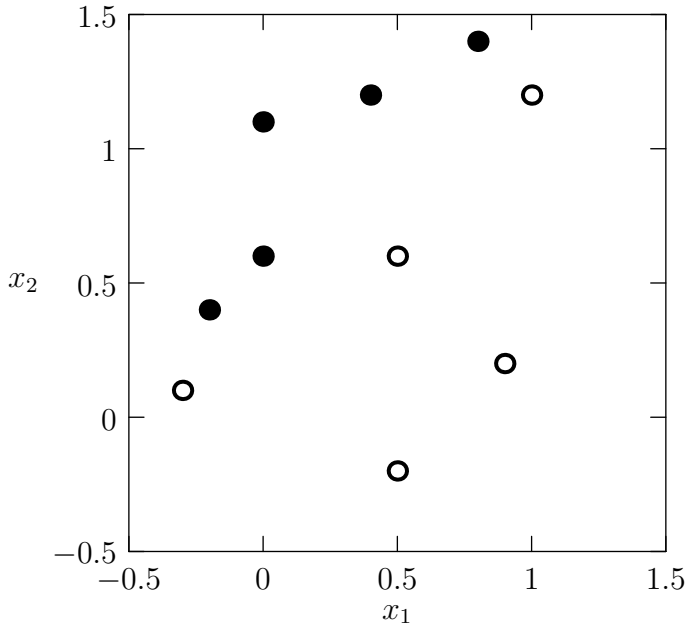
x_3 is classified as positive.

$$\text{ind}(x_4) = \text{ind}(\begin{bmatrix} 2.5 \\ 1 \end{bmatrix}) = 0.25 \cdot 4 - 0.1 \cdot 0.625 - 0.1 \cdot 0.625 = -0.25$$

x_4 is classified as negative.

b) A quadratic kernel means that the boundary must be a quadratic curve (ellipse, hyperbola or parabola). Given the datapoints from **a** means that the boundary has to pass between the pairs. Symmetry means that we also know that the boundary has to pass between $(-2, 1)$ and $(-2.5, 1)$.

B-8 Linear classifier (3p)



a) The figure illustrates a dataset where filled circles are positive samples and unfilled are negative. For a single "artificial neuron", find a set of values for the weights and the threshold that will separate the positive from the negative samples.

Note: You do not have to use any learning algorithm, but you must explain where you got the values from.

b) If a new negative sample at $(0.5, 0.0)$ arrives, and learning is done with the *perceptron learning rule*; how will these values change?

Solution:

a) A natural separator is the line passing through $(-0.5, 0.0)$ and $1.0, 1.5)$. The weight vector must be perpendicular to this line and point towards the positive sample side (upper left).

$$w_1 = -1, \qquad w_2 = 1$$

We get the threshold, θ, by testing a point on the line using the weights we have chosen.

$$-0.5 \cdot w_1 + 0.0 \cdot w_2 - \theta = 0 \quad \Rightarrow \quad \theta = 0.5$$

b) No change, because the sample is already correctly classified.