Visual Computing Institute
Computer Vision
Prof. Dr. Bastian Leibe

RWTH AACHEN UNIVERSITY

Machine Learning
SS 2015

Written Examination

2015-07-16

Name: ________________ First Name: ________________

Program of Study: ____

Matr.-No.: ________________ Exam #: ________________

Information:

- Write your name and matriculation number on **every sheet of paper**.
- Answer each question on the provided sheet. If more space is needed, **use a new sheet of paper for each question**.
- If you have to draw to answer a question, multiple templates are provided. **Cross out wrong answers!**
- At the end of the examination this cover sheet together with the question sheets and all additionally used paper has to be returned.
- Duration of the exam: **60 minutes**.
- **No additional aids** (notes, calculator, . . .) are allowed.
- Write **legibly**. Not readable text will not be graded.
- Use a pen with **blue or black ink** for writing down your solutions. Text written with pencils or red/green pens will not be graded.

With my signature I confirm that I have **read and understood** the information above.

____
Signature

Question:	1	2	3	4	5	Total
Points:	13	10	10	14	13	60
Score:						

✂ ML-Exam Number

#

Question 1: Probability Densities: . . . . ($\Sigma = 13$)

(a) For data D and hypothesis H, say whether or not the following equations must always be true. Answer in yes or no. **(2 pts)**

i. $\sum_h p(H = h \mid D = d) = 1$

i. Yes

ii. $\sum_h p(D = d \mid H = h) = 1$

ii. No

iii. $\sum_h p(D = d \mid H = h)p(H = h) = 1$

iii. No

iv. $p(H|D) = \dfrac{p(D|H)p(D)}{p(H)}$

iv. No

(b) In probability density estimation methods, we typically have a tuning parameter which acts as a smoothing factor. For example, in case of Histograms, bin size Δ is such a tuning parameter. Smoothing can also be interpreted in terms of 'bias' and 'variance'. Provide a very brief interpretation of bias and variance in terms of smoothing. How will bias and variance change if the bin size is increased? **(3 pts)**

Solution: More smoothing means high bias.
Less smoothing means high variance.
In Histograms, if bin size is increased, the resulting density will be smoother. Which also means, high bias and less variance.

(c) We want to represent the probability distribution for points x_n, $n = 1, \dots, N$ by a univariate Gaussian distribution with parameters $\theta(\mu, \sigma^2)$. Express the likelihood $p(x_n \mid \theta)$ for a single data point using the equation for the Gaussian distribution. **(1 pt)**

Solution:

$$p(x_n \mid \theta) = \mathcal{N}(x_n; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_n - \mu)^2}{2\sigma^2}\right)$$

(d) How can we estimate the parameters of θ using Maximum Likelihood for any probability distribution function? Give the main steps of the derivation. **(4 pts)**

Solution: We start by expressing the likelihood of the entire dataset

$$L(\theta) = p(X \mid \theta) = \prod_{n=1}^{N} p(x_n \mid \theta)$$

and then take the negative logarithm to obtain an energy

$$E(\theta) = -\log L(\theta) = -\sum_{n=1}^{N} \log p(x_n \mid \theta)$$

Maximizing the likelihood now corresponds to finding the minimum of the energy. In order to find this minimum, we take the derivative with respect to each parameter and set it to zero:

$$\frac{\partial}{\partial\theta} E(\theta) = -\sum_{n=1}^{N} \frac{\frac{\partial}{\partial\theta} p(x_n \mid \theta)}{p(x_n \mid \theta)} \overset{!}{=} 0$$

Solving those equations, we get the ML estimates for μ and σ^2.

(e) What implicit assumption did we make in this derivation? **(1 pt)**

Solution: We assumed that all data points are independent.

(f) What problems/limitations does Maximum Likelihood have? **(2 pts)**

Solution: ML only gives an unbiased estimate in the limit of infinite data. For small datasets, it may overfit to the given data and underestimate the variance.

Question 2: Linear Discriminant Functions: . . . . $(\Sigma = 10)$

(a) What is the difference between generative and discriminative methods for classification? **(2 pts)**

Solution: Generative methods model the joint probability distribution over observation and label sequences, whereas discriminative methods directly model the decision boundary.

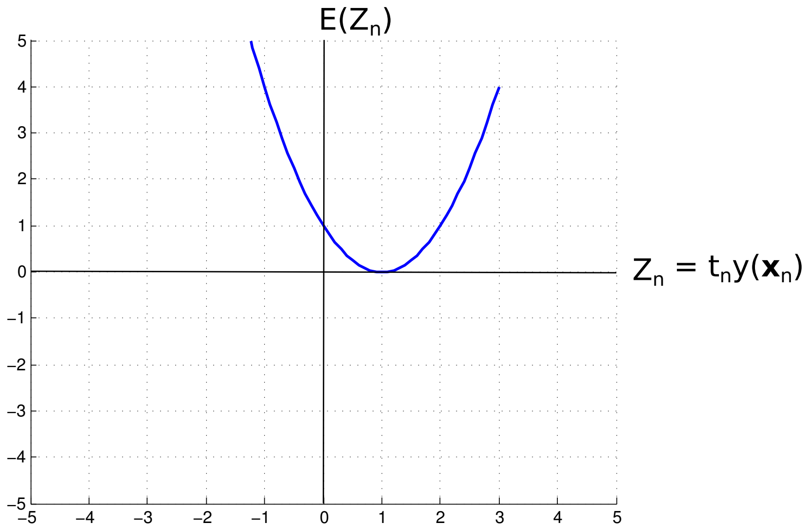
(b) Write an equation of the error function for a 2-dimensional, 2-class linear Least-Squares classifier and define the variables of the equation clearly. **(2 pts)**

Solution: For N 2D data points $\{\mathbf{x}_n\}$ and corresponding labels $\{t_n\} \in \{1, -1\}$ the least-squares error function can be written as bellow

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y(\mathbf{x_n}; \mathbf{w}) - t_n)^2$$

Where, $\mathbf{w}$ is the normal vector to the hyperplane which defines class boundary $y(\mathbf{x}) = 0$

(c) Plot the error function for the linear Least-Squares classifier. **(2 pts)**



(d) Discuss the behavior of Least-Squares classification in the presence of outliers. **(2 pts)**

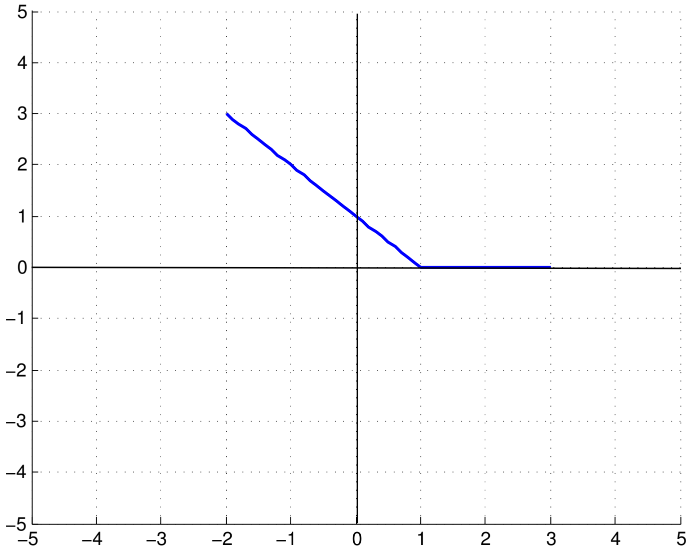
Solution: Least-squares is very sensitive to outliers. This is because the error function penalizes predictions that are “too correct”.

(e) Name two other methods that use linear discriminants and draw their corresponding error functions.

i. Name of the classification method: **(1 pt)**

i. SVM

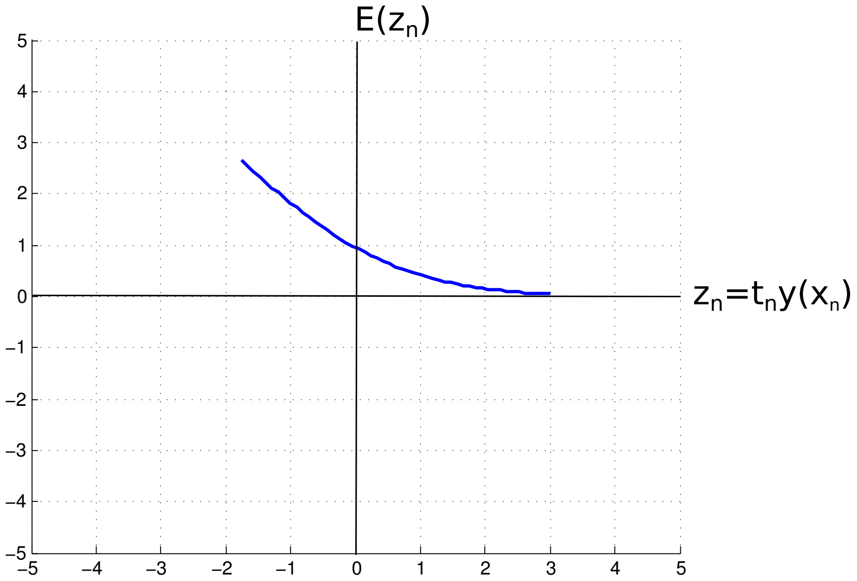
Error function:



ii. Name of the classification method: **(1 pt)**

ii. **Logistic Regression**

Error function:



Question 3: VC-dimensions and Support Vector Machines: . . . . ($\Sigma = 10$)

(a) Let H be the set of all oriented lines in the (x, y)-plane. Points on one side of the line are classified as positive and on the other side as negative. What is the VC-dimensions of H? **(1 pt)**

Solution: 3

(b) For each of the following cases, state whether it would be best to use the primal or dual SVM formulation. Briefly explain your answer. **(4 pts)**

i. We apply a feature transformation that maps the input data into a feature space with infinite dimension.

Solution: Dual.
The primal would have an infinite number of components in the weight vector $\mathbf{w}$ and be unsolvable.

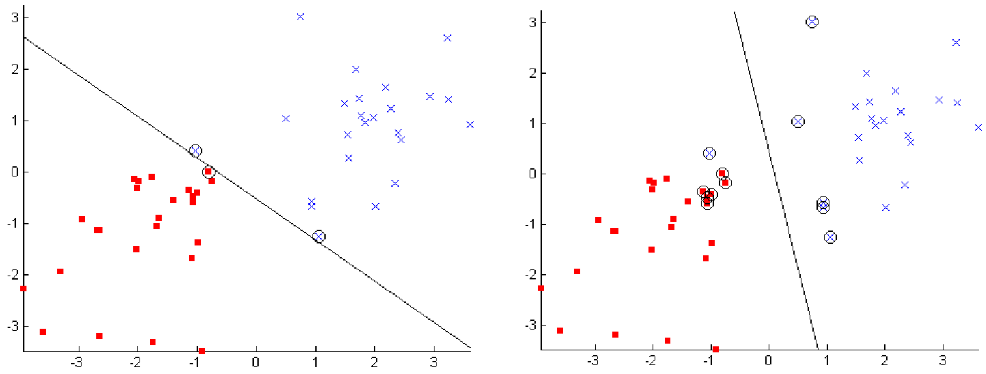
ii. We apply a feature transformation that doubles the dimension of the input data. The input data has millions of training examples and is linearly separable.

Solution: Primal.
The dual formulation would have billions of α variables, and if the data is linearly separable we do not need an parameter for each data point in the primal.

(c) Is the following statement true or false? When the data is not completely linearly separable, the linear SVM without slack variables returns $\mathbf{w} = 0$. **(1 pt)**

(c) **False**

(d) You trained a linear SVM on the toy problem below and obtained the solution shown on the left. Your friend also trained a linear SVM on the same problem, but obtained the solution shown on the right. **(2 pts)**



What happened here? How can the difference be explained?

Solution: The two solutions were generated with different values of the slack parameter C. In the left example, a large value for C was chosen, leading to a high penalty for outlier data points and therefore a very strict decision boundary. In the right image, a low value for C was used, leading to a more relaxed decision boundary.

(e) In general explain which data points will be selected as support vectors by an SVM. **(2 pts)**

Solution: The support vectors are those points that are either on the margin or on the wrong side of the margin.

Question 4: 4 . . . . **($\Sigma = 14$)**

Adaboost:

(a) Write down the steps of Adaboost algorithm below and provide the corresponding formulas. **(8 pts)**

Given a candidate pool of weak classifiers $\{h_k\}$ and training samples $\{(\mathbf{x}_n, t_n)\}$, $n = 1 \dots N$ (where $\mathbf{x}_n \in \mathbb{R}^d$ are data points in d-dimensional space and $t_n \in \{ -1, +1 \}$ are class labels), the AdaBoost algorithm for the two-class problem is:

i. Initialize the weights:

$w_n^{(1)} = \frac{1}{N}$;

ii. For $m = 1, \dots, M$

α) Train the weak classifier h_m on the weighted data; i.e., select the optimal parameter combination using the criterion $J_m = \sum\limits_{n=1}^{N} w_n^{(m)} \mathbf{I}\{t_n \neq h_m(\mathbf{x}_n)\}$

β) Compute the error: $\epsilon_m = \frac{\sum\limits_{n=1}^{N} w_n^{(m)} \mathbf{I}\{t_n \neq h_m(\mathbf{x}_n)\}}{\sum\limits_{n=1}^{N} w_n^{(m)}}$;

γ) Compute the weight for the weak classifier: $\alpha_m = \ln\left(\frac{1-\epsilon_m}{\epsilon_m}\right)$;

δ) Recalculate the weights: $w_n^{(m+1)} = w_n^{(m)} \exp\{\alpha_m t_n h_m(\mathbf{x}_n)\}$

iii. Resulting classifier:

$H(\mathbf{x}) = \operatorname{sgn}\left(\sum_{m=1}^{M} \alpha_m h_m(\mathbf{x})\right)$;

(b) Please state whether the statements below are true or false. Explain your answer.

i. AdaBoost will eventually reach zero training error, regardless of the type of weak classifier it uses, provided enough weak classifiers have been combined. **(2 pts)**

Solution: False.
Not if the data in the training set cannot be separated by a linear combination of the specific type of weak classifiers we are using.

ii. AdaBoost can model non-linear decision boundaries. **(2 pts)**

Solution: True.
If the weak decision stump is linear, the final decision boundary of the resulting strong classifier will be piecewise linear, which is effectively non-linear. If the weak decision stump is non-linear, the linear combination of them will be also non-linear.

(c) Does AdaBoost work better with strong base classifier or with weak ones? Why? **(2 pts)**

Solution: AdaBoost works better with weak base classifiers. Strong classifiers will classify a majority of the data correctly. This leaves only a small subset of misclassified data for the next iteration. In subsequent iterations, other strong classifier will then classify almost the same data points correctly. In a way, strong classifiers therefore tend to be more correlated than weak ones, which make them a poor choice for AdaBoost.

Question 5: Graphical Models . . . . ($\Sigma = 13$)

Consider the following Bayesian Network

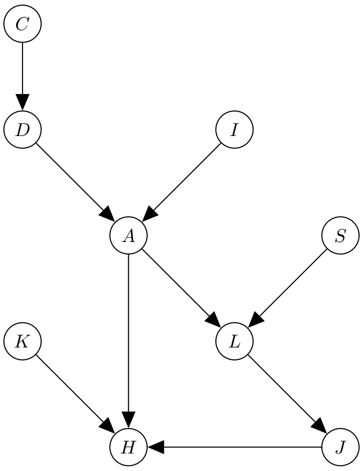


Figure 1: Bayesian Network

(a) Write down the set of variables which forms Markov blanket of 'A' in Fig. 1? **(2 pts)**

Solution:

$$\{D, I, L, S, H, K\}.$$

(b) For each of the following independence assumptions, please state whether it is true or false: **(2 pts)**

i. $D \perp\!\!\!\perp S$

i. **True**

ii. $D \perp\!\!\!\perp S \mid L$.

ii. **False**

iii. $C \perp\!\!\!\perp J \mid H$.

iii. **False**

iv. $C \perp\!\!\!\perp J \mid A$.

iv. **True**

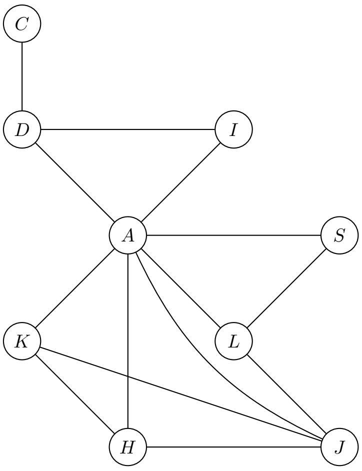
(c) Specify the factorization of the joint probability $p(A, C, D, I, S, L, J, K, H)$ of the Bayesian Network given above. **(3 pts)**

Solution:

$$\begin{aligned} p(A, C, D, I, S, L, J, K, H) = & p(C)p(D|C) \\ & p(I)p(A|D, I) \\ & p(S)p(L|A, S)p(J|L) \\ & p(K)p(H|A, J, K) \end{aligned}$$

(d) Convert the given Bayesian Network in Figure 1 into an undirected graphical model and draw the resulting graph. **(3 pts)**

Solution:



(e) Specify the factorization of the joint probability $p(A, C, D, I, S, L, J, K, H)$ of the resulting undirected model. **(3 pts)**

Solution:

$$\begin{aligned} p(A, C, D, I, S, L, J, K, H) = & \frac{1}{Z} f_1(C, D) f_2(A, D, I) \\ & f_3(A, L, S) f_4(A, L, J) \\ & f_5(A, J, H, K) \end{aligned}$$