Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Introduction to Machine Learning

Autumn Semester 2012

Prof. J.M. Buhmann

Final Exam

February 6th, 2013

First and Last name: ______________________________

ETH number: ______________________________

Signature: ______________________________

General Remarks

- You have 2 hours for the exam. There are five sections, each of which is worth 20 points. Scoring 100 points guarantees you a grade of six. In two sections you will find bonus questions, worth together 10 points. The bonus questions are a bit more difficult, we suggest you leave them to the end.
- Write your answers directly on the exam sheets. At the end of the exam you will find supplementary sheets, feel free to seperate them from the exam. If you submit the supplementary sheets, put your name and ETH number on top of each.
- Answer the questions in English. Do not use a pencil or red color pen.
- You may provide at most one valid answer per question. Invalid solutions must be canceled out clearly.

	Topic	Max. Points	Points	Signature
1	Assorted Questions	20		
2	Bayesian Inf., MAP and ML	20 + 5		
3	Supervised Learning	20		
4	Kernelized Ridge Regression	20		
5	Unsupervised Learning	20 + 5		
Total		100 + 10		

Grade: . . . .

This page has been intentionally left blank.

Question 1: Assorted Questions (20 pts.)

1. Figure 1 shows 4 times the same binary classification dataset.

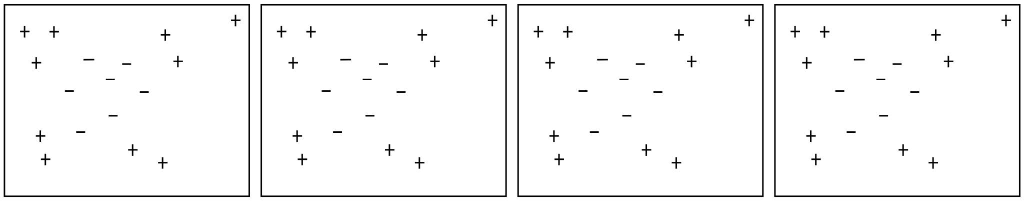


Figure 1: 4 times the same dataset

   (a) Cross all of the following algorithms/classifiers, which can achieve zero training error on this dataset.

   □ Perceptron

   ☑ Decision tree

   ☑ SVM with Gaussian kernel

   ☑ Ensemble of linear kernel SVMs

   (b) For each of the methods that can achieve zero training error, qualitatively depict a possible decision boundary (having zero error) in one of the plots of the dataset in Figure 1. Indicate which method belongs to which plot.

4 pts.

2. Let $\mathcal{F}$ be an hypothesis class for a binary classification task and f be a randomly chosen prediction function, having a training error of 0.65, on some dataset $\mathcal{S}$. Explain how to use f to obtain $\tilde{f}$, a prediction function which is **guaranteed** to have a smaller training error than f.

2 pts.

3. We consider applying the Viterbi algorithm to estimate a trajectory of an HMM with $|S|$ states over T time points. Assume that the number of states $|S|$ grows as $O(\sqrt{T})$. What is worst-case asymptotic computational complexity of the algorithm (as a function of T)?

3 pts.

4. For each of the following statements, circle the correct answer below.

   (a) The number of nodes in a decision tree is bounded by the number of features.
   True/False

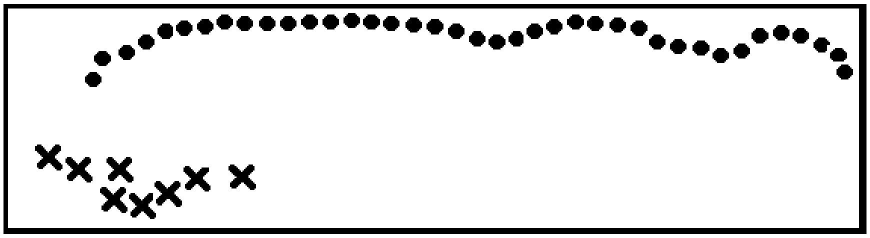
   (b) Boosting classifiers can in principle be done in a parallel manner.
   True/False

   (c) Can the Baum-Welch algorithm be considered a type of Expectation Maximization procedure?
   Always/Never/Only Sometimes

4 pts.

5. The following figures show a dataset of 48 objects from two different sources, represented by different symbols.

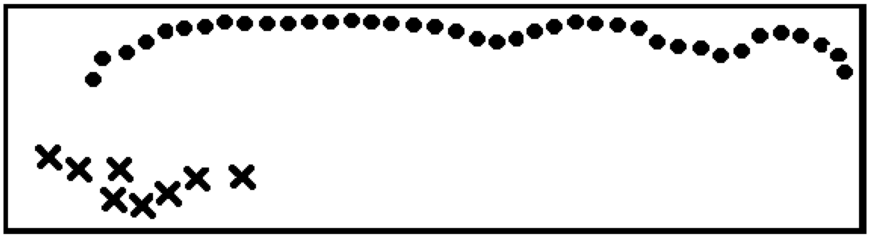
   (a) Sketch the optimal K-means solution on this dataset, for $K = 2$. Draw the centers as well as the clusters.



3 pts.

   (b) Consider reducing the dimensionality of the data to 1 before finding a 2-means solution. Propose an appropriate dimension reduction by drawing a projection line through the estimated center of mass of the data.

   Now sketch the optimal 2-means solution on the dimension reduced data.



4 pts.

Question 2: Bayesian Inference, MAP and ML estimation (20 pts.)

1. Let $x_1, \dots x_n \in \mathbb{R}$ be a dataset consisting of n samples which are assumed to be drawn iid from a normal distribution $\mathcal{N}(x|\mu, \sigma^2)$ in which the variance σ^2 and the mean μ are unknown. Demonstrate that the maximum likelihood estimation of μ can be performed without knowing the maximum likelihood estimate of the variance.

8 pts.

2. Consider the following Maximum a Posteriori estimation task. The likelihood function is the normal distribution with unknown mean μ and variance $\sigma^2 = 1$. Let μ_0 and σ_0^2 respectively denote the mean and variance of the prior, and recall that the posterior has mean and variance respectively given by

$$\begin{aligned} \mu_n &= \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right), \\ \sigma_n^2 &= \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}. \end{aligned} \tag{1}$$

(a) Show how to derive the above posterior formula for μ, from the prior and the likelihood function.

8 pts.

(b) Let $\sigma_0^2 = \pi$ and $x_i = 1$ for $i = 1, \ldots, 5$. What is the numerical value of the maximum a posteriori estimate of μ?

4 pts.

3. **Bonus question:** Let μ_{ML} and μ_{MAP} respectively denote the maximum likelihood estimator and the maximum a posteriori estimator for μ. Calculate the following:

$$\lim_{\sigma_0^2 \to \infty} \mathbb{E}[\mu_{\mathsf{MAP}}] - \mathbb{E}[\mu_{\mathsf{ML}}] \overset{?}{=}$$

5 pts.

Question 3: Supervised Learning

This question is concerned with classification of watermelons into 'good' watermelons ($+1$) and 'bad' ones (-1). Watermelons can be distinguished based only on their color and smell. Let $\mathcal{H}$ be the class of all *circles* in $\mathbb{R}^2$. We associate a classification rule with each $h \in \mathcal{H}$: the interior of the circle is classified as 'good' and outside of the circle is 'bad'.

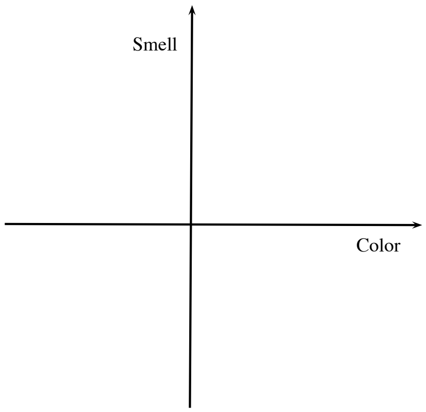
Given $\{(\mathbf{x}_i, y_i)_{i=1}^n \; \mathbf{x}_i \in \mathbb{R}^2, y_i \in \{1, -1\}\}$, a labeled sample of watermelons, we used the following criterion for the parameters of $h^* \in \mathcal{H}$:

$$w_1^*, w_2^*, r^* = \underset{w_1, w_2, r}{\operatorname{argmin}} \sum_{i=1}^{n} \exp(-y_i[r^2 - ((x_{i1} - w_1)^2 + (x_{i2} - w_2)^2)]) \tag{2}$$

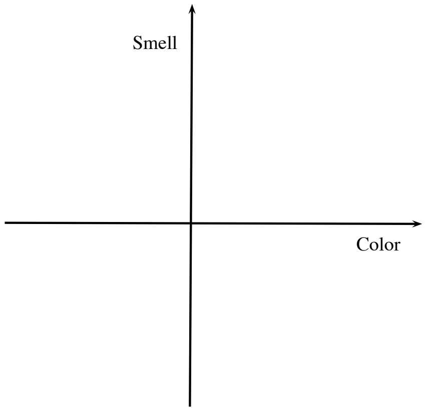
We then sold h^* to Migros as part of a watermelon test kit.
Unfortunately h^* did not meet the expectations, it misclassified a non-negligible proportion of the watermelons used at test time.

1. For each of the following additional assumptions:
   (a) Give a possible explanation for h^* performing poorly
   (b) Draw a training set, the prediction function h^*, and the true distribution (if needed) that demonstrate your explanation.

Additional assumption: h^* had a very low training error



Additional assumption: The sample size was large, and h^* had training error of ~ 0.4



8 pts.

2. Suggest a way to measure the empirical variance of the classifier h^*, given that we are out of budget for obtaining more watermelon samples.

3 pts.

Figure 2 depicts the dataset we had, and h^* that we got using equation (2). To improve h^* we decided to add a regularizing term, the new criterion will be

$$\underset{w_1,w_2,r}{\operatorname{argmin}} \sum_{i=1}^{n} \exp(-y_i[r^2-((x_{i1}-w_1)^2+(x_{i2}-w_2)^2)])+\lambda\Omega(w_1,w_2,r) \tag{3}$$

Where $\Omega(w_1, w_2, r)$ is the regularizer. We ask you to suggest a suitable regularizer.

4. (a) Draw on Figure 2 the regularized solution you envision.

   (b) Write down the mathematical term of the regularizer. Explain your answer.

$\Omega(w_1, w_2, r) =$

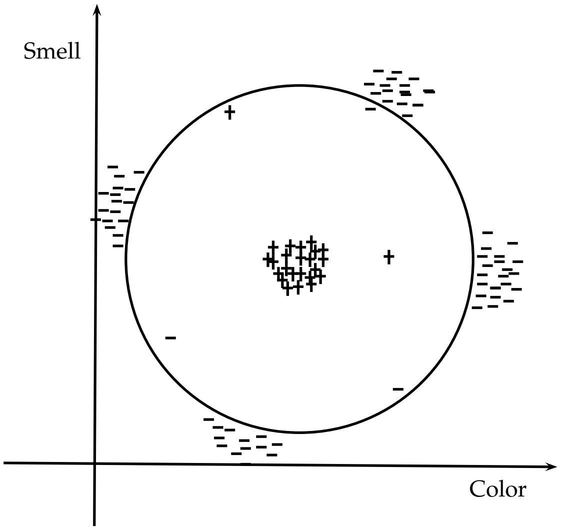


Figure 2: Watermelons dataset and h^*

5 pts.

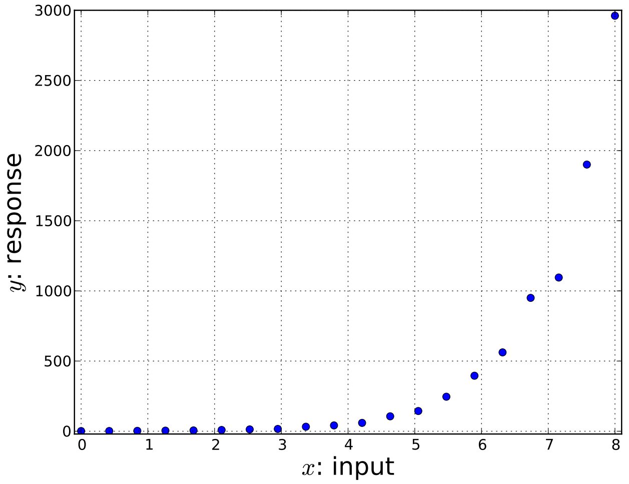
5. Assume that the true distribution of watermelons consists of high density regions visible in Figure 2, plus sparse outliers. Explain what happens to the variance of h^* as we increase λ compared to some starting value $\lambda_0 > 0$.

4 pts.

Question 4: Kernelized Ridge Regression (20 pts.)

Recall the regression setting: Given input vectors $\mathbf{x}_i$, and output (response) variables y_i, the goal is to find a functional relation between them, often expressed with a weight vector $\mathbf{w}$ and bias b.

1. Below is a dataset with one dimensional input variables x, and response variable y. Your task is to find a kernel function $K(x_i, x_j)$, such that you can use a linear regression method in the kernel space.



   $K(x_i, x_j) = \dots$ **3 pts.**

2. You will now derive a kernelized version of ridge regression by introducing a feature transform $\mathbf{\Phi}(\mathbf{x}_i)$. This should allow a non-linear regression solution, for datasets such as the one depicted above.
   Recall the formulation of ridge regression as an optimization problem:

$$\min_{\mathbf{w},b} \quad \sum_i (y_i - \mathbf{w}^\top \mathbf{x}_i - b)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 \tag{4}$$

   (a) We replace the inputs $\mathbf{x}_i$ in Equation (4) with the vectors of the features in the kernel space and rewrite the problem as a constrained optimization problem by introducing the new variables ξ_i. Write down the equality constraint in Equation (6).

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \sum_i \xi_i^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 && (5) \\ \text{s.t.} \quad & \xi_i = \ldots && (6) \end{aligned}$$

2 pts.

(b) Write down the Lagrangian of this new optimization problem using $\boldsymbol{\alpha}$ as the dual variable.

(c) Derive the dual optimization problem.

8 pts.

3. Express the dual problem in terms of the kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$.

3 pts.

4. Given the optimal solution of the dual problem $\boldsymbol{\alpha}^\star$ and a new point $\mathbf{x}_k$, write down the equation to compute y_k.

4 pts.

Question 5: Unsupervised Learning (20 pts.)

1. In this section we study non-parametric density estimation of an arbitrary point x. We consider some small region $\mathcal{R}$ containing x. In the class we have seen the following generic formula for density estimation:

$$p(x) = \frac{K}{nV} \ ,$$

where K denotes the number of data points falling inside the region $\mathcal{R}$ and V shows the volume of the region. n is the number of data points in the sample set $\mathcal{S} = \{x_1, ..., x_n\}$.

(a) Consider the following Gaussian distribution to be used as a Parzen window function:

$$\phi(x - x_j) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{(x - x_j)^2}{2}\right) . \qquad (7)$$

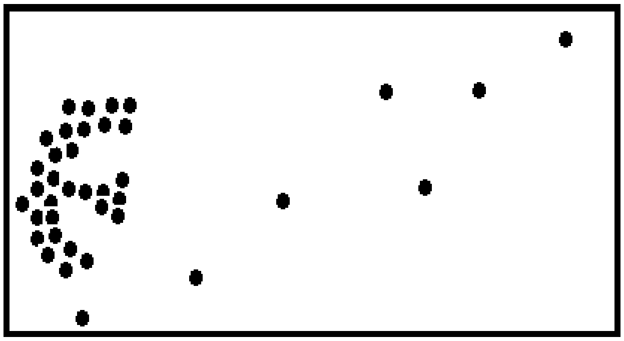
What are K and V for this window function?

4 pts.

(b) This particular choice of a window function leads to underfitting. Add a parameter to increase the model complexity.

2 pts.

(c) Consider the following sample set. Which of the density estimation methods would you choose? Window-based (Eq. (7)) or K-nearest neighbor? Explain your answer.



3 pts.

(d) For a general Parzen window function prove that it provides a probability distribution.

4 pts.

2. We consider a mixture of K poisson distributions and perform the Expectation-Maximization (EM) algorithm to compute the unknown parameters. The log-likelihood function of n independent objects for mixture of K Poisson distribution is defined as:

$$P(x; \lambda) = \sum_{i=1}^{n} \log \sum_{c=1}^{K} \pi_c f(x_i; \lambda_c)$$

where π_c's are the mixture weights and λ_c's are the parameters of K Poisson distributions. $f(x_i; \lambda_c)$ is defined as:

$$f(x; \lambda_c) = \frac{\lambda^x e^{-\lambda_c}}{x!}.$$

(a) Introduce the latent indicator variables necessary for maximizing the log-likelihood function.

2 pts.

(b) Calculate the expectation of the latent variables. Provide a Bayesian interpretation for your answer.

5 pts.

(c) **Bonus question:** Assume the expectations of the latent variables are given. Calculate the unknown parameters λ_c's. Write down the details of your calculations.

5 pts.

Supplementary Sheet

Supplementary Sheet

Supplementary Sheet

Supplementary Sheet