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: ....

Question 1: Assorted Questions (20 pts.)

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

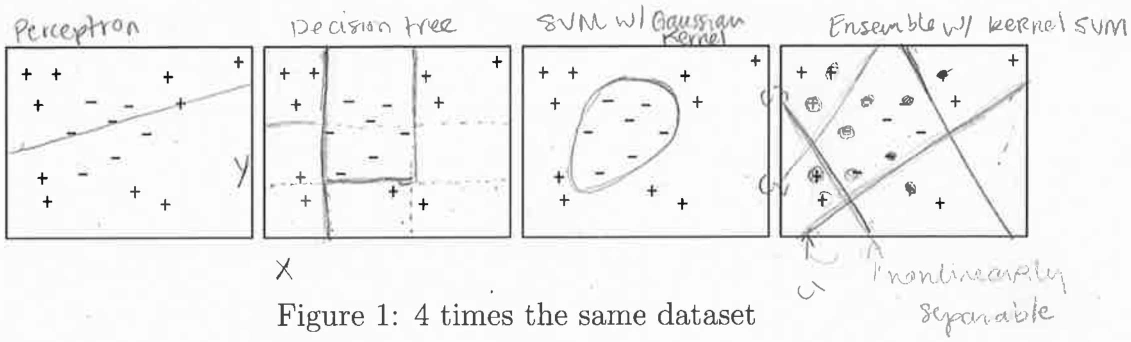


Figure 1: 4 times the same dataset

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

- ☐ Perceptron   no data needs to be linearly separable.
- ☒ 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.

predictions   0/1

f : prediction fxn = Train error 0.65

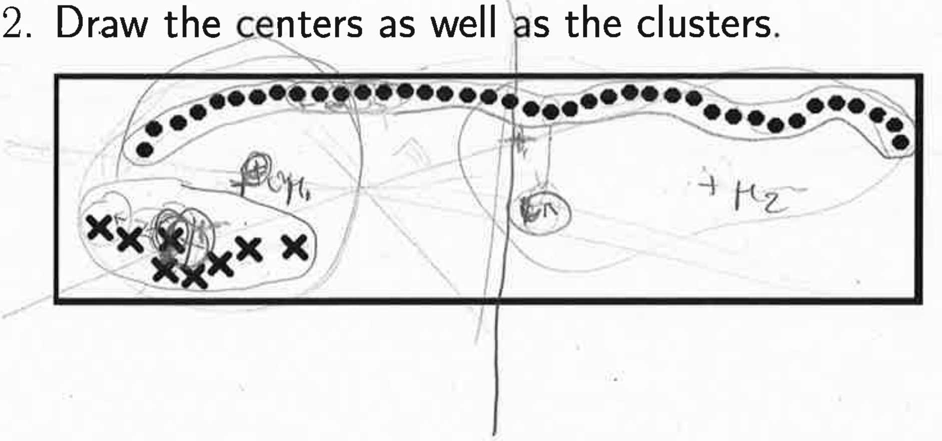
$f_R \to \hat{f}_{pred}$ regularized form of f and perform CV.

2 pts.

f can be evaluated by selecting a cost fxn and performing a cross validation on the number of predicted fxns to choose the best parameters.
OR in ensemble methods update the weight of each classifier such as boosting.

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

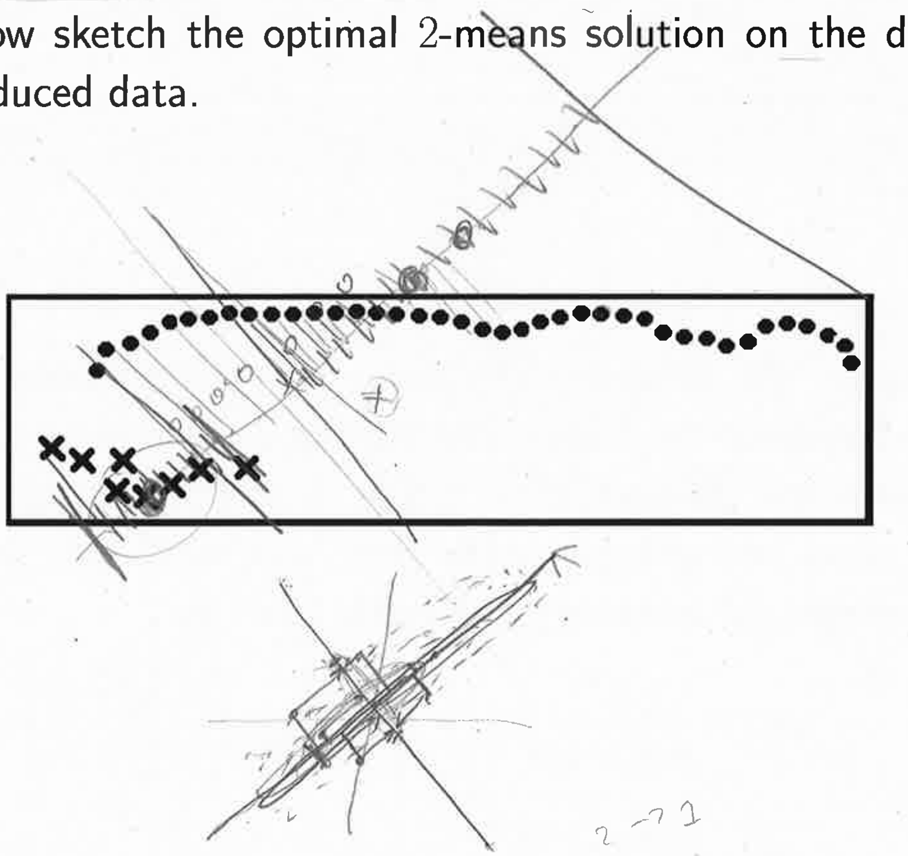
(a) Sketch the optimal K-means solution on this dataset, for $K = 2$. Draw the centers as well as the clusters.



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.

2 → 1

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

Def'n of MAP

$$\hat{\mu}_{MAP} = \arg\max_{\mu} \underline{P(x_1, \dots, x_n \mid \mu, \sigma)\, p(\mu \mid \mu_0, \sigma_0)}$$

PRIOR

$\div p(x_1, \dots, x_n \mid \mu_0, \sigma_0)$ ← can throw away because doesn't depend on μ.

MAP is a linear combination of ML

$\hat{\mu}_{MAP} =$ scale accordingly $(\hat{\mu}_{ML})$ + constant

$$= \arg\max_{\mu} p(x_1, \dots, x_n \mid \mu, \sigma) \cdot p(\mu \mid \mu_0, \sigma_0)$$

= multiplying 2 gaussians together.

likelihood

$$\arg\max_{\mu} \left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2} \right) \right] \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right)$$

scaling doesn't affect max / $\hat{\mu}$ value

$$= \arg\max_{\mu} \exp\left(-\frac{1}{2} \left[\left[\sum_{i}^{N} \frac{(x_i - \mu)^2}{\sigma^2} \right] + \frac{(\mu - \mu_0)^2}{\sigma_0^2} \right] \right) = \arg\min_{\mu} \left[\sum_{i}^{N} \frac{(x_i - \mu)^2}{\sigma^2} \right] + \frac{(\mu - \mu_0)^2}{\sigma_0^2}$$

$\hat{\mu}_{MAP}$ such that $\frac{\partial}{\partial \hat{\mu}_{map}} \left(\left[\sum_{i}^{n} \frac{(x_i - \mu)^2}{\sigma^2} \right] + \frac{(\hat{\mu}_{ML} - \mu_0)^2}{\sigma_0^2} \right) = 0$

$$= 2 \left[\sum_{i}^{N} - \frac{x_i}{\sigma^2} - \frac{\mu_0}{\sigma_0^2} + \hat{\mu}_{MAP} \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \right] = 0 \Rightarrow \hat{\mu}_{MAP} = \frac{n\sigma_0^2}{\frac{n}{\sigma^2} + \frac{1}{\sigma^2}} \left(\frac{1}{n} \sum x_i \right) + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0$$

8 pts.

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

$$\hat{\mu}_{MAP} = \frac{5\pi + \mu_0}{5\pi + 1}$$

4 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 $h^* \in \mathcal{H}$:

cost   classifiers

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

center   radius

Trained classifier

$-y_i (r^2 - (x_i - w_i)^2)$

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.

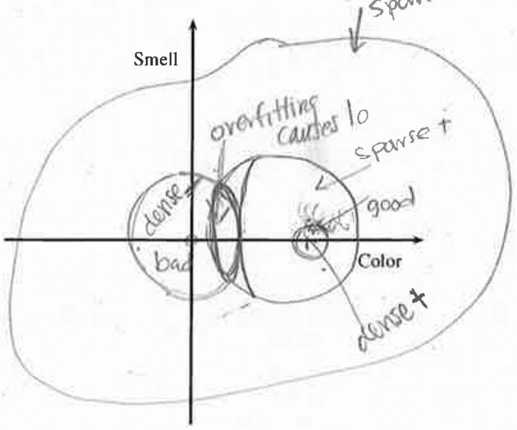
① overfitted h^*

Know that the distribution of the watermelons is not 50/50 and you have to consider the

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

additional assumption that getting 105/39 and give you a dense solution for good

Added noise to true distribution



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

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

$\Omega(w_1, w_2, r) =$ R^2 in L2-norm.

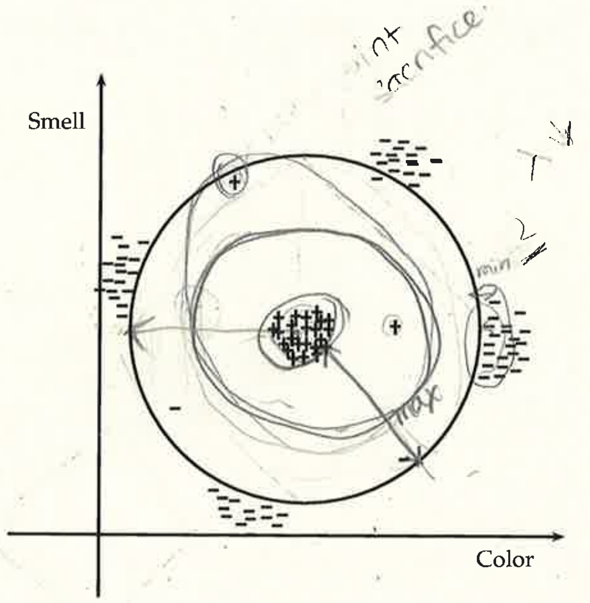


Figure 2: Watermelons dataset and h^*

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 $\lambda_0 > 0$.

The size the circle will decrease.

variance ↓ 150

4 pts.

constrained optimization problem by introducing the new variables ξ_i. Write down the equality constraint in Equation (6).

primal

$$\min_{\mathbf{w},b,\xi} \sum_i \xi_i^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 \tag{5}$$

$$\text{s.t.} \qquad \xi_i = \ldots \tag{6}$$

2 pts.

(b) Write down the Lagrangian of this new optimization problem using α as the dual variable.

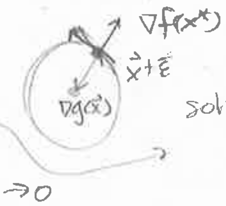
objective of lagrangian is to minimize the problem.

$\max \ f(\vec{x})$

s.t. $g(\vec{x})=0 \qquad L(\vec{x},\lambda) = f(\vec{x}) + \lambda g(\vec{x})$

$\vec{x} \in \mathbb{R}^D$

$\nabla g(\vec{x}) \perp \quad g(\vec{x}+\vec{\varepsilon}) \simeq g(\vec{x}) + \vec{\varepsilon}^T \nabla g(x)^T \quad \|\vec{\varepsilon}\| \to 0$



solve the larangian

$\frac{\partial L}{\partial x} = 0$

$\frac{\partial L}{\partial \lambda} = g(x) = 0$

(c) Derive the dual optimization problem.

objective fxn

$L(\vec{\alpha}, \vec{w}, b, \vec{\xi}) = \sum_i \xi_i^2 + \frac{\lambda}{2}\|\vec{w}\|^2 + \sum_i \alpha_i (y_i - \vec{w}^T \vec{\phi}_i - b - \xi_i)$

$\vec{w} = \binom{b}{\vec{w}} \quad \vec{x} = \binom{1}{\vec{x}}$

$\frac{\partial L}{\partial \vec{w}} = -\sum_i \alpha_i \vec{\phi}_i + \lambda \vec{w} = 0 \Rightarrow \boxed{\vec{w} = \frac{\sum_i \alpha_i \vec{\phi}_i}{\lambda}}$ This is why you derive on α.

$\frac{\partial L}{\partial \xi_i} = 2\xi_i - \alpha_i = 0 \Rightarrow \boxed{\xi_i = \frac{\alpha_i}{2}}$

$\|\vec{w}\|^2 = \frac{\lambda}{2} \cdot \frac{1}{\lambda^2} \|\sum_i \alpha_i \vec{\phi}_i\|^2 = \sum_{i,j} \alpha_i \alpha_j \vec{\phi}_i^T \vec{\phi}_j$ You want to get the objective fxn on alpha.

$L(\vec{\alpha}) = -\frac{1}{4}\sum_i \alpha_i^2 + \sum_i \alpha_i y_i - \frac{1}{2\lambda}\sum_{i,j} \alpha_i \alpha_j \vec{\phi}_i^T \vec{\phi}_j, \quad \min_{\vec{\alpha}} L(\vec{\alpha})$

8 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:

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

What are K and V for this window function?

$K = \sum_{j=1}^{n} \phi(x - x_j)$

$V = 1$ ← Because a probability distribution

4 pts.

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

Add variance parameter. $= h$, controls smoothness

$\phi\left(\frac{(x-x_j)^2}{h}\right) = \frac{1}{(2\pi)^{\frac{1}{2}} \cdot h} \exp\left(-\frac{(x-x_j)^2}{2h^2}\right)$

2 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:

$1 \cdot k \cdot K$

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

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

$\log P(x|\vec{\lambda},\vec{\pi}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k f(x_n|\lambda_k)$

$Z \in \mathbb{R}^{N \times K}$, each row is a vector for a point

$Z_{nk} = \begin{cases} 1 : \text{component } k \text{ is responsible for generated } x_n \\ 0 : \text{o.w.} \end{cases}$

2 pts.

(b) Calculate the $\underbrace{\text{expectation of the latent variables}}_{\text{Conditional Expectation}}$. Provide a Bayesian interpretation for your answer.

E-step: estimate posterior distribution of data

$p(z|x,\theta^{old}) = E_z(z|\dots)$

M-step: $\theta^{new} = \arg\max_{\theta} Q(\theta,\theta^{old})$

$Q(\theta,\theta^{old}) = \sum_{z} P(z|x,\theta^{old}) \cdot \log P(X,Z|\theta)$; complete data log-likelihood

joint distribution

$\underbrace{E[Z_{nk}]}_{\text{Responsibility } \lambda(Z_{nk})} = P(Z_{nk}=1|\vec{x}_n) * 1 = \dfrac{\boxed{\underbrace{P(Z_{nk}=1)}_{=\pi_k} \cdot \underbrace{P(\vec{x}_n|Z_{nk}=1)}_{=P(x_n|\lambda_k)}}}{\boxed{\sum_{Z_1} P(X_n, Z_n)}} = \sum_{j=1}^{k} \pi_j \cdot f(x_n|\lambda_j)$

$P(\vec{Z}_n|\vec{\pi}) = \prod_{k=1}^{K} \pi_k^{Z_{nk}}$

$\Rightarrow P(x_n|Z_n,\lambda_k) = \prod_{k=1}^{K} p(x_n|\lambda_k)^{Z_{nk}}$

5 pts.

Supplementary Sheet

NO 'fixed "right solution", as in real life. → The accuracy is the only goal.

Google solution for insights and do it better → do not reinvent the wheel.

CV scheme is really important

SVM → assume metric space behind data.

Q/A session before exam.

Supplementary Sheet