ETH
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.

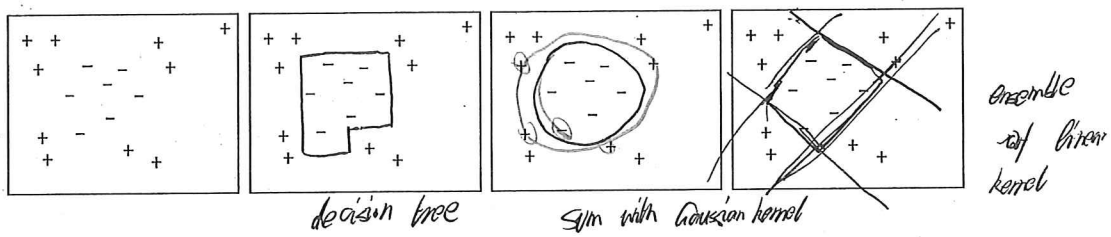


Figure 1: 4 times the same dataset

(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.

use the ensemble methods — adaboost,

for $b = 1 : B$:

train $G_b(x)$ on $w_i^{(b-1)}$ weighted dataset $\mathcal{S}$

compute $\epsilon_b = \frac{\sum_{i=1} w_i^{(b-1)} \mathbb{I}(G_b(x_i) \neq y_i)}{\sum_{i=1} w_i^{(b-1)}}$

compute $\alpha_b = \frac{1}{2} \log \frac{1-\epsilon_b}{\epsilon_b}$

reweight $w_i^{(b)} = w_i^{(b-1)} e^{\alpha_b \mathbb{I}(G_b(x_i) \neq y_i)}$

$G(x) = \sum_{b=1}^{B} \alpha_b G_b(x)$

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

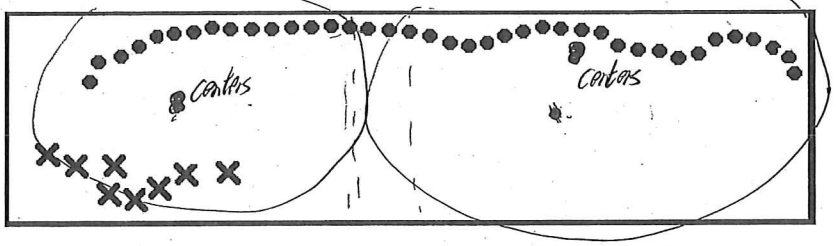
(c) Can the Baum-Welch algorithm be considered a type of Expectation Maximization procedure?
Always/Never/Only Sometimes

是 EM算法在HMM的具体表现

4 pts.

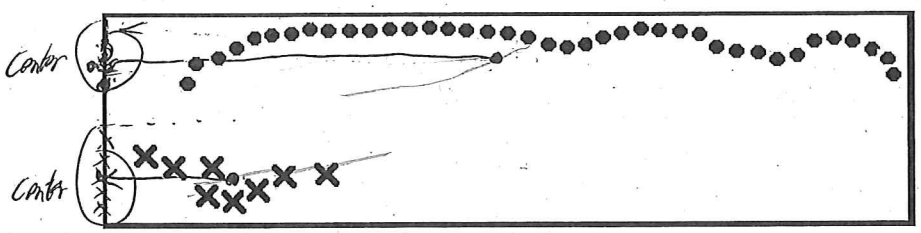
5. The following figures show a dataset of 48 objects from two different sources, represented by different symbols.

   (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.

$$L = \sum_{i=1}^{n} \log P(x|\mu,\sigma^2)$$

$$= \sum_{i=1}^{n} \left(-\frac{(x_i-\mu)^2}{2\sigma^2} - \frac{1}{2}\log 2\pi - \log \sigma \right)$$

$$\frac{\partial L}{\partial \mu} = \sum_{i=1}^{n} \frac{2(x_i-\mu)}{2\sigma^2} = 0$$

$$\Rightarrow \sum_{i=1}^{n} (x_i-\mu) = 0$$

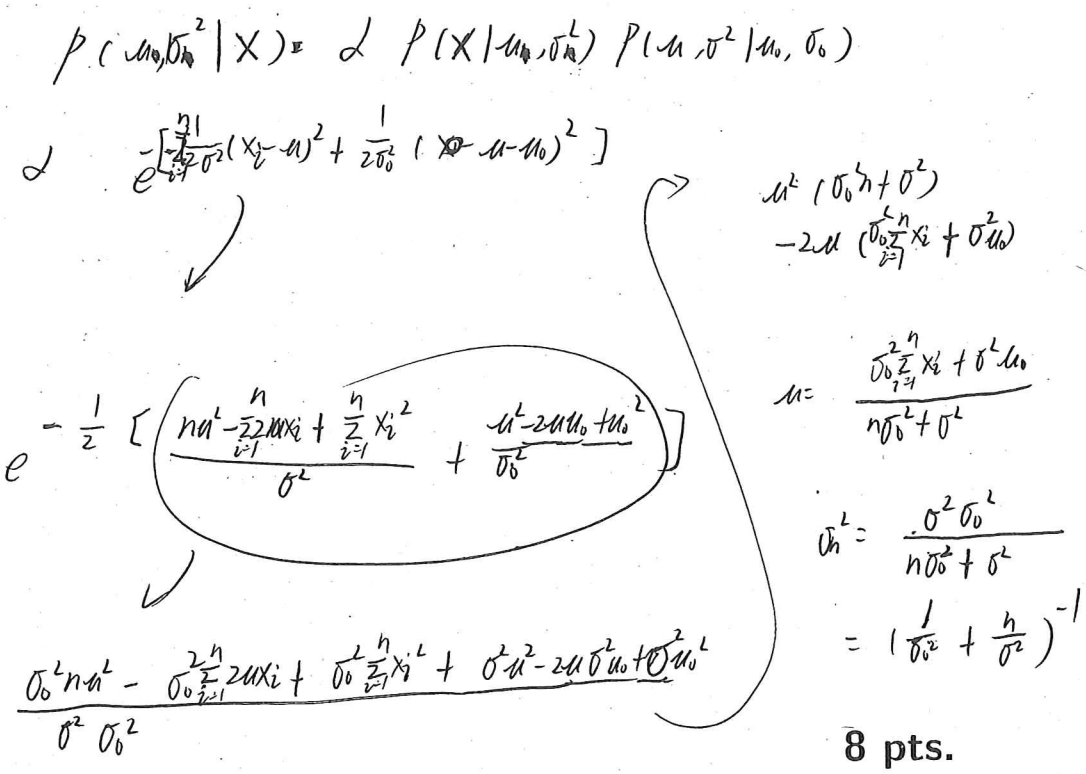
$$\Rightarrow \mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

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?

$$\mu_n = \frac{1}{5\times\pi + 1}\mu_0 + \frac{5\times\pi}{5\times\pi + 1}\left(\frac{1}{5}\sum_{i=1}^{n} x_i\right)$$

$$= \frac{\mu_0}{5\times\pi + 1} + \frac{5\pi}{5\pi + 1}$$

$$= \frac{5\pi + \mu_0}{5\pi + 1}$$

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:

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

$$\mathbb{E}\left[\frac{\sum_{i=1}^{n} x_i}{n}\right]$$

$$\mathbb{E}\left[\lim_{\sigma_0^2 \to \infty} \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) - \frac{\sum x_i}{n}\right]$$

$$= 0$$

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'.

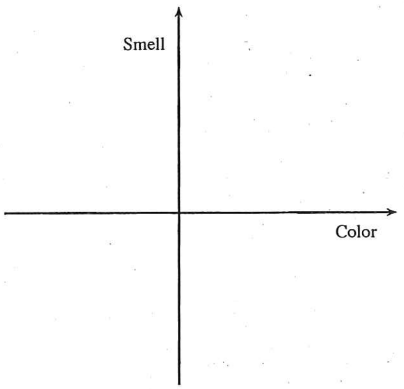
Given $\{(\mathbf{x}_i, y_i)_{i=1}^n \; \mathbf{x}_i \in \mathbb{R}^2, y_i \in \{1, -1\}\}$, a labeled sample of watermelons, we used the following criterion for the parameters of $h^* \in \mathcal{H}$:

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

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.

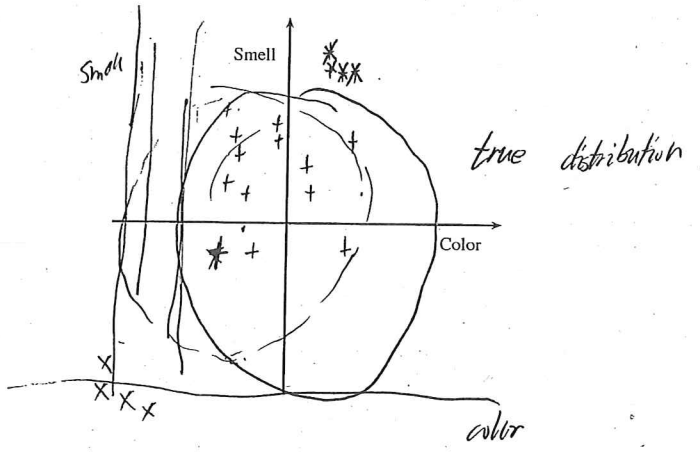
1. For each of the following additional assumptions:
   (a) Give a possible explanation for h^* performing poorly
   (b) Draw a training set, the prediction function h^*, and the true distribution (if needed) that demonstrate your explanation.

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



(a) overfitting, the model too complex

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.

use cross-validation K-fold

for $k = 1 : K$

train $C_k(x)$ based on the $n\frac{k-1}{K}$ dataset

the empirical variance

$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(y_i \neq C_{k(i)}(x_i))$

$\frac{1}{n}\sum_{i=1}^{n} (y_i - C_{k(i)}(x_i))^2$

$x_i \in$ $k(i)$-th fold dataset

$\frac{1}{k}\sum_{k=1}^{K}\left[(w_{1k} - \frac{k}{\sum_{i=1} w_{1k}}{K})^2 + (w_{2k} - \frac{k}{\sum_{i=1} w_{2k}}{K})^2 + (r_k - \frac{k}{\sum_{i=1} r_k}{K})^2\right]$

$\frac{1}{k}\sum_{k=1}^{K}\left(C_{k(i)} - \frac{k}{\sum_{i=1} C_{k(i)}}{K}\right)^2$

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.

$(w_1 - w_{10})^2 + (w_2 - w_{20})^2 + (r - r_0)^2$

where w_{10}, w_{20}, r_0 are the prior knowledge we have about the shape of the boundary

4. (a) Draw on Figure 2 the regularized solution you envision.
   (b) Write down the mathematical term of the regularizer. Explain your answer.

$\Omega(w_1, w_2, r) =$ $(w_1 - w_{10})^2 + (w_2 - w_{20})^2 + (r - r_0)^2$

we assume r_0 not that large thus penalize the way to make r in the decision boundry large. we also make prior assumption about the typical characteristic of a good water melon. thus give the w_{10} and w_{20}

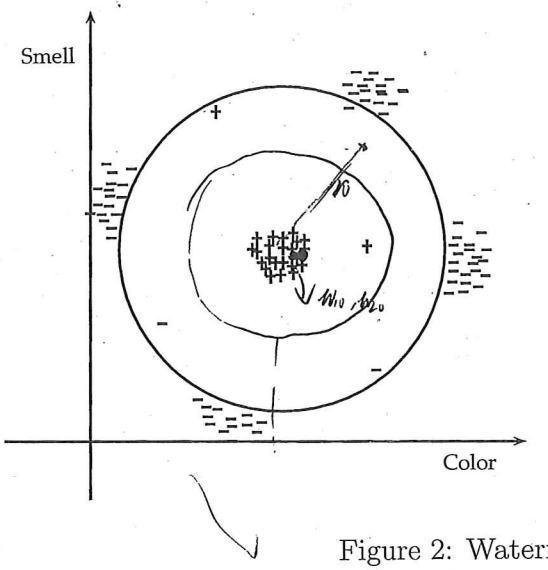


Figure 2: Watermelons dataset and h^*

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 $\lambda_0 > 0$.

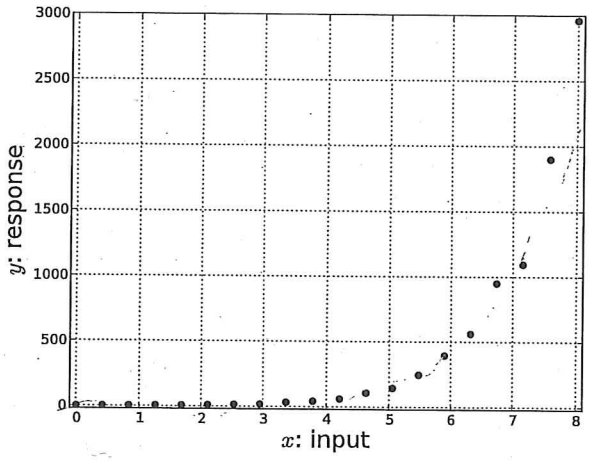
the variance become smaller, it give negative incentive for ~~while the bias be a~~ h^* to be a large circle. thus make r smaller and close to r_0 at the same time give constrains to the w_1 and w_2, make it ~~close~~ to a smaller distance to w_{10}, w_{20} of the

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.



$e^{x_1} e^{x_2}$

$K(x_i, x_j) = \ldots$ ~~$(e^{x_i}+1)(e^{x_j}+1)$~~ $e^{(x_i+x_j)}$ **3 pts.**

2. You will now derive a kernelized version of ridge regression by introducing a feature transform $\Phi(\mathbf{x}_i)$. 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).

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

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

$y_i - w^T \phi(x_i) - b$

2 pts.

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

$L(\text{Lagrangian}) = \sum_i \rho_i^2 + \frac{\lambda}{2}\|w\|^2 + \sum_i \alpha_i (\rho_i - y_i + w^T \phi(x_i) + b)$

(c) Derive the dual optimization problem.

$$\frac{\partial L}{\partial w} = \lambda w + \sum_i \alpha_i \phi(x_i) = 0$$

$$w = \frac{-\sum_i \alpha_i \phi(x_i)}{\lambda}$$

$$\frac{\partial L}{\partial b} = \sum_i \alpha_i = 0$$

$$\frac{\partial L}{\partial \rho_i} = 2\rho_i + \alpha_i = 0$$

$\alpha_i = -2\rho_i \qquad \rho_i = -\frac{\alpha_i}{2}$

$$L = \sum_i \frac{\alpha_i^2}{4} + \frac{\lambda}{2} \frac{\|\sum_i \alpha_i \phi(x_i)\|^2}{\lambda^2} + \sum_i \alpha_i \left(-\frac{\alpha_i}{2} - y_i + \frac{-\sum_j \alpha_j \phi(x_j)\phi(x_i)}{\lambda}\right)$$

$$= \sum_i \frac{\alpha_i^2}{4} + \frac{\sum_{ij} \alpha_i \alpha_j \phi(x_i)\phi(x_j)}{2\lambda} + -\frac{\sum_i \alpha_i^2}{2} - \sum_i \alpha_i y_i - \frac{\sum_{ij} \alpha_i \alpha_j \phi(x_i)\phi(x_j)}{\lambda}$$

$$= -\frac{\sum_i \alpha_i^2}{4} - \frac{\sum_{ij} \alpha_i \alpha_j \phi(x_i)\phi(x_j)}{2\lambda} - \sum_i \alpha_i y_i$$

maximize $W(\alpha) = -\frac{\sum_i \alpha_i^2}{4} - \frac{\sum_{ij} \alpha_i \alpha_j \phi(x_i)\phi(x_j)}{2\lambda} - \sum_i \alpha_i y_i$

with $\sum_i \alpha_i = 0$

8 pts.

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

$$W(\alpha) = -\frac{\sum \alpha_i}{4} - \frac{\sum \alpha_i \alpha_j y_i y_j K(x_i, x_j)}{\lambda} - \sum \alpha_i y_i$$

subject to $\sum_i \alpha_i y_i = 0$

3 pts.

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

$$y_k = w^T \phi(x_k) + b$$

$$b = \frac{1}{n} \sum_i \left[\frac{\sum_i \alpha_i^* y_i \phi(x_i) \phi(x_j)}{\lambda} - y_j \right]$$

$$w^* = \frac{-\sum_i \alpha_i^* y_i \phi(x_i)}{\lambda}$$

4 pts.

$$y_k = \frac{-\sum_i \alpha_i^* y_i K(x_i, x_k)}{\lambda}$$

when conditional number ratio of the largest eigenvalue to $(X^TX)^{-1}X^TY$ to the least eigenvalue is large

f) Model inference (computing model parameters) for the *RSS* cost function requires the inversion of a matrix: $(\mathbf{X}^\top\mathbf{X})^{-1}$.
- Specify a mathematical condition when this inversion is *numerically unstable.* the range of different feature is too large
- Describe in your own words under which circumstance this instability happens during a practical application of regression.
- Are we then in risk of *under-* or *over-fitting*?
- Also comment qualitatively on the bias and variance of the model parameter's estimation in this circumstance.

X^TX is not positive matrix $\forall a \; a^TX^TXa > 0$ not persist.

3.5 pts.

over-fitting

in that case, the

variance will be ~~huge~~ big

the bias will be small

no regularization.

when the input data sample are correlated highly

feature large sample small

in the extream ~~ext~~ case,

when the data X sample are the same

which means $X_1 \cdots X_N$ are the same

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \end{pmatrix}$$

$$(y_i - \sum_{i=1}^{101} x_i \beta_i) = 0$$

we can find β_i infinite number of β_i who meet this situation.

feature selection

g) We can stabilize the inversion by reducing the model complexity.
- Explain **how** *Ridge Regression* (*RR*) limits the model complexity. the $\lambda\beta^T\beta$ factor somehow represents the prior knowledge of the distribution for β, to be $N(0, \Lambda^{-1})$ ..
- Provide another approach to limiting the model complexity. add constraint for β, just like LASSO.
- Demonstrate the stabilization mathematically by writing down the *Ridge Regression* solution and argue with the Eigenvalues of the matrix that needs to be inverted.

feature selection

4 pts.

in RR the β would be $(X^TX+\lambda I)^{-1}X^TY$

can be express by $\sum_{j=1}^{d} u_j \frac{d_j^2}{d_j^2+\lambda} u_j^T y$

in this case, we add the shrinkage factor which mean the small value will be balance by λ. thus adding stability and omit the ..

$$X\hat{\beta} = X(X^TX+\lambda I)^{-1}X^TY$$
$$= UD(D^2+\lambda I)^{-1}DU^TY$$
$$= \sum_{j=1}^{d} u_j \frac{d_j^2}{d_j^2+\lambda} u_j^T Y$$

h) Comment qualitatively on the bias and variance of *Ridge Regression* as compared to the *RSS* estimator.

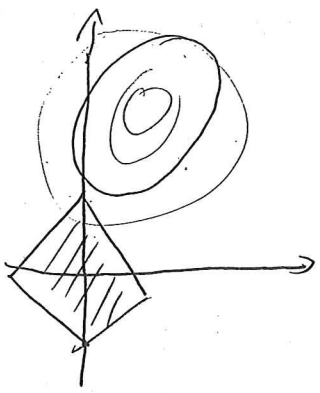
1.5 pts.

the bias ~~become large~~ is bigger

while the variance decreases

i) By depicting an appropriate plot (including notation), provide a graphical argument why *LASSO* favors sparse solutions.

1 pt.



Question 5: Unsupervised Learning (20 pts.)

a) For each of the following non-parametric approaches mention the most important parameter that influences the smoothness of the results: **3 pts.**

- histograms

  # of bins (the length of bins)

- Parzen window estimates

  size of the h_n

- nearest neighbor estimates.

  number of k

b) Determine whether the following statements are true or false. Briefly explain your answer. **2 pts.**

- Non-parametric estimation methods are less sensitive than parametric approaches to model misspecification.

  Yes, since non-parametric way does not specify an distribution in advance, putting more emphasis on data

- In histograms, by changing the dimensionality the number of required bins (to keep the resolution) increases linearly with dimension.

  no, polynomially, d is the dimension

  $(N)^d$

c) Consider Hidden Markov Models (HMMs). For each of the following algorithms determine if it solves a supervised or an unsupervised problem. Explain your answer.

4 pts.

- Viterbi algorithm

unsupervised learning we only observe the emitted sequence since we know the evidence variable while the hidden path is unknown

- Baum-Welch algorithm.

unsupervised

d) In this section we study the k-means clustering method.

1. Mention at least two main differences between k-means and Gaussian Mixture Model (GMM) clustering methods.

2 pts.

k-means doesn't assume the underlying distribution of the data while gmm assume the model of the distribution, which is gaussian - not only give the classification, by also give the probability of the specific point assigned to different cluster.

Consider the k-means cost function defined as:

$$R^{km} = \sum_{n=1}^{N} \sum_{l=1}^{k} r_{nl} ||\mathbf{x}_n - \boldsymbol{\mu}_l||^2. \tag{2}$$

Here $\boldsymbol{\mu}_l$ denotes the l-th centroid and $r_{nl} \in \{0, 1\}$ indicates the assignment of object $\mathbf{x}_n$ to the l-th cluster.

2. Write down the assignment update step (E-step) for the k-means algorithm.

2 pts.

$$C^{(t)}(x_i) = \underset{c(x) \in \{1,\dots k\}}{\operatorname{argmin}} \|x - \mu_{c(x)}\|^2$$

3. Derive the centroids update step (M-step). We expect you to write down all intermediate steps of the centroid update derivation.

4 pts.

$$D = \sum_i^{h_\alpha} (x_i - \mu_\alpha)^2 = 0$$

$$\frac{\partial D}{\partial \mu_\alpha} = -2 \sum_{i}^{h_\alpha} x_i - \mu_\alpha = 0$$

then $\mu_\alpha = \frac{1}{n_\alpha} \sum_{x:c(x)\in\alpha} x$ with $n_\alpha = \#\{x : c(x) \in \alpha\}$

until change to $c(x)$, γ vanish.

4. Show that the k-means algorithm always converges.

3 pts.

Since there are only finite way of assign sample to k groups

every iteration, the configuration changes and distortion improves,

Supplementary Sheet