Prof. J.M. Buhmann

Final Exam

February 8th, 2012

First and Last name: ______________________

ETH number: ______________________

Signature: ______________________

	Topic	Max. Points	Points	Signature
1	Bayesian Inference and MLE	20		
2	Linear Classifiers and Kernels	20		
3	Ensemble Methods	20		
4	Regression, Bias and Variance	20		
5	Unsupervised Learning	20		
Total		100		

Grade: . . . .

Question 1: Bayesian Inference and Maximum Likelihood(20 pts.)

A telecommunication company needs to estimate the rate of the telephone calls in a small town in order to adjust its channel capacity. We model a length of a time-interval between telephone calls as a random variable x. Knowing that a sequence of calls is the realization of the Poisson process, we model the time it takes before the next call using the exponential distribution.

Consider the task of estimating the rate parameter λ of the exponential distribution from n i.i.d. observations $\mathcal{X} = \{x_i\}_{i=1}^n$, $x \in \mathbb{R}$

$$Exp\left(x|\lambda\right) = \lambda \exp\left(-\lambda x\right).$$

a) Write the maximum likelihood estimator for the rate $\widehat{\lambda}_{\mathsf{ML}}(\mathcal{X})$ as an explicit function of the i.i.d. observations $\mathcal{X} = \{x_i\}_{i=1}^n$: (please write the direct closed-form solution and the derivation)

1. $\widehat{\lambda}_{\mathsf{ML}}(\mathcal{X}) = \arg\max p(\mathcal{X}|\lambda) =$ **2 pts.**

2. Where did you use the fact that the observations are i.i.d? **1 pt.**

To find the posterior density $p(\lambda|\mathcal{X})$ we need a prior on λ. We claim that a conjugate prior for the exponential distribution is the gamma distribution

$$Gamma\,(\lambda|\alpha,\beta) = \frac{\Gamma(\alpha)}{\beta^{\alpha}}\lambda^{\alpha-1}\exp(-\lambda\beta),$$

where $\Gamma(\alpha) = \int_0^\infty \exp^{-t} t^{\alpha-1}dt$ is the gamma function.

b) 1. What does *conjugate prior* mean? **1 pt.**

2. Show that the gamma distribution is the conjugate prior of the exponential distribution. **2 pts.**

c) Given a Gamma prior over the rate λ (prior with parametes α and β), write the maximum a posteriori estimator $\widehat{\lambda}_{\mathsf{MAP}}(\mathcal{X})$ as an explicit function of the i.i.d. observations $\mathcal{X} = \{x_i\}_{i=1}^n$: (please write the direct closed-form solution)

$\widehat{\lambda}_{\mathsf{MAP}}(\mathcal{X}) = \arg\max p(\lambda|\mathcal{X}) =$ **2 pts.**

d) When is the maximum likelihood estimatior (MLE) equal to the maximum a posteriori (MAP) estimator given a set of i.i.d. observations $\mathcal{X} = \{x_i\}_{i=1}^{n}$?

1. If the number of observations is finite. **2 pts.**

2. If the number of observations is infinite $(n \to \infty)$. **2 pts.**

e) Assume that you have a set of the i.i.d. observations $\mathcal{X} = \{x_i\}_{i=1}^{n}$ and that you can not decide which distribution to use for data description: the Gaussian distribution or the Beta distribution.

If you use the Bayesian framework what you can look at? **2 pts.**

Now consider a binary classification task from a set of the i.i.d. observations $\mathcal{X} = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$, with $\mathbf{x} \in \mathbb{R}^D$. Assume that the likelihood of both classes is Gaussian (assume class prior π_i, mean μ_i, and covariance matrix Σ_i for class y_i, with $i = 1, 2$).

f) Recall that a discriminant function for class y_i is defined as:

$$g_{y_i}(\mathbf{x}) = p(y_i|\mathbf{x}) .$$

How can you find a decision surface in terms of likelihood, prior and evidence? **1 pt.**

g) Assume that

$$\begin{aligned}
&\mu_1 = \mu_2 = \mu \\
&\Sigma_1 = \tfrac{1}{2\lambda_1}\mathbb{I}, \quad \Sigma_2 = \tfrac{1}{2\lambda_2}\mathbb{I} \\
&\lambda_1 > 0, \quad \lambda_2 > 0, \quad \lambda_1 \neq \lambda_2
\end{aligned}$$

where $\mathbb{I}$ denotes the identity matrix. Write the equation satisfied by the separating decision surface. The equation must be an explicit function of $\mathbf{x}_1$ (the single observation), of the class prior, means and covariance:
(please write the solution in the polinomial form)

1. Decision surface: **3 pts.**

2. In the case described above, is the decision surface linear, parabolic, spherical, cylindrical, or something else?

   linear|parabolic|spherical|cylindrical|other **2 pts.**

Question 2: Linear Classifiers and Kernels (20 pts.)

a) Below is a list of algorithms which given a training set output a prediction function. Cross **all** of the algorithms that necessarily output a linear (in the original space) prediction function.

□ Neural network
□ Perceptron with learning rate $\eta = 1$
□ SVM with radial basis kernel
□ K-nearest neighbor classifier
□ SVM with polynomial kernel with degree 1
□ Ridge regression

3 pts.

b) Recall the SVM problem. As a constrained optimization problem a solution can be obtained through both the primal and the dual form.

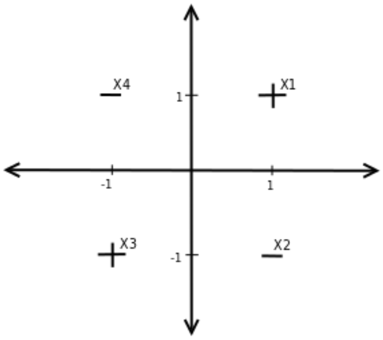
1. Given a primal solution for the SVM, write down the resulting classifier. **1 pt.**

2. Given a dual solution for the SVM, write down the resulting classifier. **1 pt.**

3. In practice, often the dual form of the SVM is solved to obtain a classifier. Provide one advantage of solving the dual SVM instead of the primal. **3 pts.**

c) Let $S = \{\mathbf{x}_i, y_i\}_{i=1}^{4}$ be the following training set

$$\begin{aligned} \mathbf{x}_1 &= (1,1) & y_1 &= 1, \\ \mathbf{x}_2 &= (1,-1) & y_2 &= -1, \\ \mathbf{x}_3 &= (-1,-1) & y_3 &= 1, \\ \mathbf{x}_4 &= (-1,1) & y_4 &= -1 \end{aligned}$$



Suppose that we trained an SVM on S and the resulting classifier $f(x)$ achieved zero training error. We ask you to provide an explicit description of $f(x)$ (a formula with numeric values).
Hint: Think of a suitable kernel function or alternatively a feature map. **5 pts.**

d) Consider a training set $S = \{\mathbf{x}_i, y_i\}_{i=1}^n$ where $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \{-1, 1\}$

1. Briefly describe a leave one out (LOO) procedure for estimating the error of an SVM classifier on S. **2 pts**

2. What is the LOO error? **1 pts**

3. Suppose that we trained an SVM classifier on the **entire** dataset S, denote by sv the set of support vectors, $sv = \{\mathbf{x}_j | \alpha_j > 0\}$.
   For the same value of C, prove that the LOO error is bounded by $\frac{|sv|}{n}$ i.e.

$$\text{LOO error} \leq \frac{|sv|}{n}$$

4 pts.

Question 3: Bagging and Boosting (20 pts.)

a) Answer precisely the following questions.

1. Are bagging and Boosting Bayesian approaches? Why?

   1 pt.

2. How is it possible to detect outliers with AdaBoost?

   1 pt.

3. From the frequentist perspective, bagging is motivated by the tradeoff between two terms. Which?

   1 pt.

4. AdaBoost has an alternative interpretation which is based on the minimization of a certain cost function. Which function?

   1 pt.

5. AdaBoost aims at selecting the best approximation to which ratio?

   2 pt.

6. Why is the standard form of AdaBoost limited to binary classification?

1 pt.

7. How could one parallelize bagging?

1 pt.

8. Name a design property of the base classifiers of AdaBoost which impacts the overall predictive power.

1 pt.

9. Under which conditions does AdaBoost yield good results even when the base classifiers exhibit an individual performance that is only slightly better than that purely due to chance?

2 pt.

b) Consider *bagging* in the context of binary classification. Let the *target function* be $h(x)$, where $h : \mathbb{R}^d \to \{\pm 1\}$. Let us combine B *individual classifiers* $y_b(x), b = 1 \dots B$ to obtain a *committee model*

$$y_{\mathsf{COM}}(x) = \text{sign}\left[\frac{1}{B}\sum_{b=1}^{B} y_b(x)\right]. \tag{1}$$

1. Write down the pseudocode of bagging for binary classification, from the input (data and model) to the prediction output.

3 pts.

2. The error $\epsilon_b(x) = \exp\{-h(x)y_b(x)\}$ indicates the error of an individual model $y_b(x)$ for a single sample x in terms of the target function $h(x)$ and the output of the individual model $y_b(x)$. Write down E_{AV}, that is the average of the expected errors over the individual classifiers $y_b(x)$, and the expected error E_{COM} made by combined model $y_{\mathsf{COM}}(x)$ as a function of the output of the committee model and of the target function.

$E_{\mathsf{AV}} =$

$E_{\mathsf{COM}} =$

1 pts.

3. Under which conditions is $E_{\mathsf{AV}} < E_{\mathsf{COM}}$?

3 pts.

4. With the same exponential error, write down the error function for each iteration of AdaBoost with weighting coefficients α_b for the B base classifers $y_b(x)$.

   $E_{\mathsf{AdaBoost}} =$

1 pt.

5. In this scenario, within AdaBoost the minimization of this error function is performed with respect to two terms, which?

   1)

   2)

1 pt.

Question 4: Regression, Bias and Variance (20 pts.)

a) Write down the linear regression model (component-wise and in vector notation) for input variable $x = (1, x_1, \ldots, x_D)^\mathsf{T} \in \mathbb{R}^{D+1}$ and output variable y. Formally introduce the model parameter(s).

From now on, assume that the input dataset consists of N samples given by the matrix $\mathbf{X} \in \mathbb{R}^{N \times (D+1)}$ (where the first column is $\mathbf{1}$), and the output, $\mathbf{Y} \in \mathbb{R}^{N}$. Write down the linear regression model for all observations $\mathbf{Y}$ (in matrix notation).

2 pts.

b) Write down the following cost functions (in a notation of your choice): **2 pts.**

1. Ridge Regression (*RR*):

2. Least Absolute Shrinkage and Selection Operator (*LASSO*). Formulate as a constrained optimization problem:

c) Formulate the objective of **learning** (formula) in the regression setting introduced above with data $(\mathbf{X}, \mathbf{Y})$ i.i.d. from $P(\mathbf{X}, \mathbf{Y})$.

2 pts.

d) Assuming that the observations in $\mathbf{Y}$ are affected by additive Gaussian noise ϵ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. What do we know about the distribution of the *RSS*-estimator (i.e. $\hat{\beta}^{RSS}$)?

3 pts.

e) Please briefly give another motivation for the *RSS* cost function.

1 pt.

f) Model inference (computing model parameters) for the *RSS* cost function requires the inversion of a matrix: $(\mathbf{X}^{\top}\mathbf{X})^{\text{-}\mathbf{1}}$.

- Specify a mathematical condition when this inversion is *numerically unstable*.

- 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.

3.5 pts.

g) We can stabilize the inversion by reducing the model complexity.
- Explain **how** *Ridge Regression* (*RR*) limits the model complexity.
- Provide another approach to limiting the model complexity.
- Demonstrate the stabilization mathematically by writing down the *Ridge Regression* solution and argue with the Eigenvalues of the matrix that needs to be inverted.

4 pts.

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

1.5 pts.

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

- Parzen window estimates

- nearest neighbor estimates.

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.

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

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

- Baum-Welch algorithm.

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.

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.

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

4 pts.

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

3 pts.

Supplementary Sheet

Supplementary Sheet

Supplementary Sheet