Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Introduction to Machine Learning

Autumn Semester 2011

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 $\mathcal{X} = \{x_i\}_{i=1}^n$, $x \in \mathbb{R}$

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

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. $\widehat{\lambda}_{\mathsf{ML}}(\mathcal{X}) = \arg\max p(\mathcal{X}|\lambda) = \prod_{i=1}^{n} \lambda \exp(-\lambda x_i)$ **2 pts.**

$\Rightarrow \log\left(\sum_i^n \lambda \exp(-\lambda x_i)\right)$

$\sum_i^n (\log(\lambda) - \lambda x_i) = n \log \lambda - \lambda \sum_i^n x_i \stackrel{!}{=} 0$

$\frac{d}{d\lambda}\left(n \log \lambda - \lambda \sum_i^n x_i\right) \stackrel{!}{=} 0 \quad n\frac{1}{\lambda} - \sum_i^n x_i = 0 \quad \frac{n}{\lambda} = \sum_i^n x_i$

$$\boxed{\hat{\lambda}_{ML} = \frac{n}{\sum_i^n x_i}}$$

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

$\prod_{i=1}^{n}$ in the likelihood equation

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

$p(\lambda|x) = p(x|\lambda)\, p(\lambda)$ ← old distribution

↑ distribution you assumed that x is drawn from

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

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.**

the prior distribution $P(\lambda)$ is conjugate to the likelihood distribution $p(x(\lambda)$ if multiplying these two distributions together and normalizing results in another distribution of the same form as the prior $P(\lambda)$.

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

exp × gamma

$= \left(\lambda \exp(-\lambda x)\right)\left(\frac{\Gamma(\alpha)}{\beta^\alpha}\lambda^{\alpha-1}\exp(-\lambda\beta)\right) = \lambda^1 \cdot \lambda^{\alpha-1} \exp^{-\lambda x - \lambda\beta}$

$= \lambda^{1+\alpha-1} \exp^{-\lambda(x+\beta)}$

$\boxed{= \lambda^{(\alpha+1)-1} \exp^{-\lambda(x+\beta)}}$

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

$\widehat{\lambda}_{\text{MAP}}(\mathcal{X}) = \arg\max p(\lambda|\mathcal{X}) = \prod_{i=1}^{n}\left[\left(\lambda\exp(-\lambda x)\right)\left(\frac{\Gamma(\alpha)}{\beta^\alpha}\lambda^{\alpha-1}\exp(-\lambda\beta)\right)\right]$ **2 pts.**

$= \prod_{i=1}^{n}\left[\lambda^{\alpha}\exp^{-\lambda(x_i+\beta)}\right] \xrightarrow{\log} \sum_i^n \left(\log(\lambda^\alpha) - \lambda(x_i-\beta)\right)\frac{d}{d\lambda} \stackrel{!}{=} 0 \quad n \cdot \frac{1}{\lambda^\alpha} - \sum_i^n (x_i - \beta) = 0$

$\frac{1}{\lambda^\alpha} = \frac{\sum_i^n (x_i-\beta)}{n} \Rightarrow \lambda^{\alpha \cdot \frac{1}{\alpha}} = \left(\frac{n}{\sum_i^n x_i - \beta}\right)^{\frac{1}{\alpha}}$

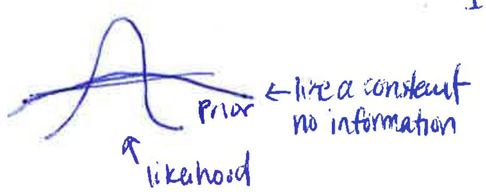
$\boxed{\cancel{\lambda_{\text{MAP}} \left(\frac{n}{\sum_i^n x_i - \beta}\right)^{\frac{1}{\alpha}}}}$

$\lambda_{map} = \frac{\alpha - 1 + n}{\beta + \sum x_i}$

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.**

NO. 1 sample



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

yes Always converges

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.**

conjugate prior and likelihood

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 $i = 1, 2$).

multivariate normal: $\frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})\right)$

$|\Sigma| \Rightarrow$ det Σ

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

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

class decision: $g_y(x) > g_z(x)$ $\forall z \neq y$

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

$p(y_i|x) = p(x|y_i)p(y_i) =$

$$p(y_i|x) = \frac{\left(\prod_{i=1}^{n} \frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)\right)\right) \cdot \pi_i}{\int_{-\infty}^{\infty} \left(p(x|y_i) \cdot \pi_i \right) dx_i}$$

1.5 Sl 27

Decision surface $g_1 = g_2$

4

$p(y=1|x)\pi_1 > p(y=2|x)\pi_2$

Decision discriminant fxn = posterior, defines boundary if one is greater

$\rightarrow \frac{p(x|y_1)p(y_1)}{p(x)} > \frac{p(x|y_2)p(y_2)}{p(x)}$

g) Assume that

$$\mu_1 = \mu_2 = \mu$$
$$\Sigma_1 = \frac{1}{2\lambda_1}\mathbb{I}, \quad \Sigma_2 = \frac{1}{2\lambda_2}\mathbb{I}$$
$$\lambda_1 > 0, \quad \lambda_2 > 0, \quad \lambda_1 \neq \lambda_2$$

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: p 36
(please write the solution in the polinomial form)

$g_1(x) - g_2(x) = 0$

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

$\ln(P(x_1 | \omega_1)) + \ln P(\omega_1) - \ln(P(x|\omega_2)) - \ln(P(\omega_1))$

$- \frac{1}{2} x^T W x + \bar{w}^T x + w_{i0}$ $\quad -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma|$

$W = \Sigma^{-1}$ $\quad w = \Sigma^{-1}\mu$ $\quad \frac{1}{2}\mu^T\Sigma^{-1}\mu$

$\Sigma_1 = \Sigma_2 \to$ linear $\quad \Sigma_1 \neq \Sigma_2$.

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

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

Dependent on covariance matrix

Gaussian within a gaussian.

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.

inputs

output

step fn

- ☒ 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 $\vec{w}\,\phi(x)$

also works w/ kernels. **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.**

   $\cancel{W = \sum_i \alpha_i y_i x_i}$ $\cancel{w^T x + w_0 = }$

   $y = \text{sgn}(w^T x + w_0)$

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

   $W = \sum_i \alpha_i y_i x_i \rightarrow w^T x + w_0 = y.$

   $\alpha_i \geq 0$

   $y = \text{sgn}\left(\sum_i \alpha_i y_i x_i \, X + w_0\right)$

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.**

Even though the primal form it has been transformed into an.

$y = w^T x + w_0$ $\max_\alpha \left(\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j \right)$

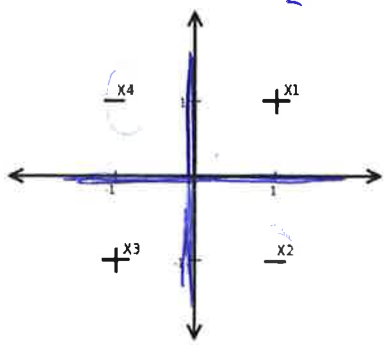
$k(x,y) = \langle \phi(x), \phi(y) \rangle$

$w = \sum \alpha_i y_i x_i^T x_i$

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

$\min \frac{1}{2}\|w\|^2 + \alpha(\cdots)$

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



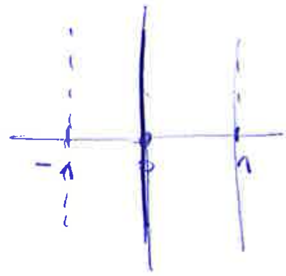
Suppose that we trained an SVM on S and the resulting classifier $f(x)$ achieved zero training error. We ask you to provide an explicit description of $f(x)$ (a formula with numeric values).
Hint: Think of a suitable kernel function or alternatively a feature map. **5 pts.**

$\mathrm{sgn}(\langle a, b \rangle)$ scalar product

$\phi(x) = \langle \phi(x_1), \phi(x_2) \rangle$

$\prod_i x_i \prod_j x_j$

$K(x_i, x_j) =$



$k(x,y) = \prod_i x_i \prod_j x_j$ $= k(x_1) = 1 \cdot 1 = 1$

$k(x_2) = 1 \cdot -1 = -1$

$k(x_3) = -1 \cdot -1 = 1$

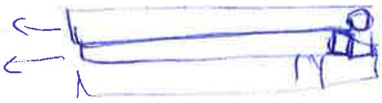
$k(x_4) = -1 \cdot 1 = -1$

Closure property:

$K(x,y) = f(x) \cdot f(z)$

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**



$\mathbb{I}$

$R^{cv} = \frac{1}{n} \sum_n \mathbb{I}$ indicator fxn

$\frac{\text{number of errors}}{10}$

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

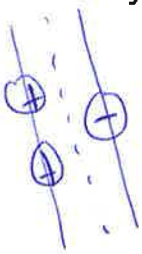
$\Rightarrow R^{cv} = \frac{1}{n} \sum_n \mathbb{I}_{y_i \neq c^{\setminus K(i)}(x(i))}$.

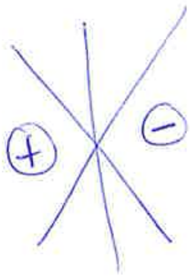
3. Suppose that we trained an SVM classifier on the **entire** dataset S, denote by sv the set of support vectors, $sv = \{\mathbf{x}_j | \alpha_j > 0\}$.
For the same value of C, prove that the LOO error is bounded by $\frac{|sv|}{n}$ i.e.

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

cardinality → number of sv

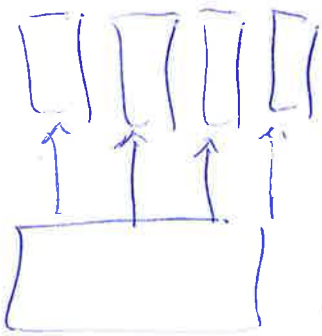
4 pts.





$R^{cv} = \frac{1}{n} \sum_{i=1}^{n} R^{-\mu}$

$= \frac{1}{n} \sum_{\mu \in sv} R^{-\mu} \leq \frac{1}{n} |sv| \cdot c$



Question 3: Bagging and Boosting (20 pts.)

a) Answer precisely the following questions.

1. Are bagging and Boosting Bayesian approaches? Why?

Not Bayesian approaches. → pg. 410 Hastie **1 pt.**

655 Bagging construct new parameters ← approximate parameters prior

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

Super large weights at the end of the training to one classifier will detect **1 pt.**

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

Bias, variance **1 pt.**

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

exponential loss **1 pt.**

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

log-odds ratio **2 pt.**

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

1 pt.

$\hat{C_b} = \mathrm{sgn}\left(\sum_{i}^{B} \alpha_b C_b(x)\right)$

7. How could one parallelize bagging?

subsets

1 pt.

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

uncorrelated
different

weights to classifiers.

1 pt.

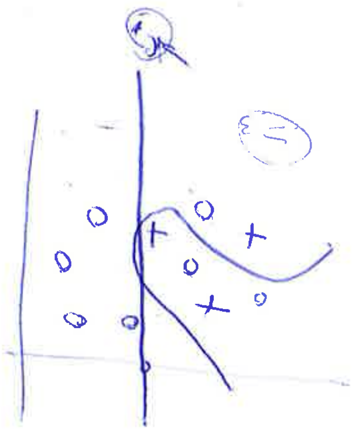
weakness of basic learners
strong and weak learners.

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?

when uncorrelated to different classifiers

2 pt.

classifiers give different outputs
and answers.
B is much bigger than 1.
and much larger dataset



Supplementary Sheet

Supplementary Sheet