ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Introduction to Machine Learning

Autumn Semester 2013

Prof. Andreas Krause

Final Exam

February 3rd, 2014

First and Last name: _____

ETH number: _____

Signature: _____

General Remarks

- You have 2 hours for the exam. There are seven sections, each of which is worth 17 points.
- 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	Naive Bayes	17		
2	Support Vector Machines	17		
3	Kernels	17		
4	Probabilistic Modeling	17		
5	Gaussian Processes	17		
6	Boosting	17		
7	Clustering	17		
Total		119		

Grade: . . . . . .

Question 1: Naive Bayes (17 pts.)

1. Consider the problem of three-class classification with labels $y \in \{-1, 0, +1\}$. Assume that the prior over the labels has the form: $P(Y = 1) = p$, $P(Y = -1) = q$, $P(Y = 0) = 1 - p - q$, s.t. $p, q \in [0, 1]$, $p + q \leq 1$. Estimating the parameters of the prior over labels $P(y)$ is the first step necessary in Naive Bayesian Classification. Assume there are n observed labels $\mathbf{y} = \{y_1, \ldots, y_n\}$.

   (a) Assume there are n_+ positive-labelled data points, n_- negative-labelled data points and n_0 zero-labelled data points ($n_+ + n_- + n_0 = n$). Write down the likelihood function of the labels given the parameters and explain, what assumption is made about the labels of the data points:
   $P(\mathbf{y} \mid p, q) =$

   ? pts.

   (b) **Derive** the MLE estimators $\hat{p}$, $\hat{q}$, given $n_+ = 5$, $n_- = 4$, $n_0 = 1$.
   Hint: Write down the maximization/minimization problem and solve it.

? pts.

2. In this question we consider Naive Bayesian Classification for two-class (labels $y \in \{-1, +1\}$) classification with Gaussian features. Assume that the following functional dependencies hold for a vector of features (x_1, x_2, x_3):

$$x_1 = z,\ x_2 = 2z,\ x_3 = t + 3,$$

where $P(z|y = +1) = \mathcal{N}(\mu_+, 1)$, $P(z|y = -1) = \mathcal{N}(\mu_-, 1)$, $P(t|y = +1) = \mathcal{N}(\nu_+, 1)$ and $P(t|y = -1) = \mathcal{N}(\nu_-, 1)$. Also, z is independent of t.

You are also given the prior $P(y = +1) = P(y = -1) = 0.5$

(a) Write down the likelihood $p(x_1, x_2, x_3|y)$:

? pts.

(b) Write down the MLE estimators for μ_+, μ_-, ν_+, ν_- given the two observations: $(x_1, x_2, x_3|y) = (1, 2, 6|+1), (0, 0, 3|-1)\}$ (no derivation needed, only general formula and numeric value).

? pts.

(c) Given a new data-point $\mathbf{x}' = (2, 4, 5)$ calculate the posterior probability $P(y|\mathbf{x}')$.

? pts.

(d) Derive the decision rule using Bayesian decision theory and give the prediction for $\mathbf{x}'$.

? pts.

Question 2: Support Vector Regression (17 pts.)

The cost function for support vector regression (SVR) is given by

$$\text{minimize } \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}(\xi_{i,+} + \xi_{i,-})$$

$$\begin{aligned} \text{s.t. } \quad y_i - \mathbf{w}^T\mathbf{x}_i &\leqslant \epsilon + \xi_{i,+} \\ \mathbf{w}^T\mathbf{x}_i - y_i &\leqslant \epsilon + \xi_{i,-} \\ \xi_{i,+}, \xi_{i,-} &\geqslant 0; i = 1, \ldots, n \end{aligned} \tag{1}$$

where training data is given by data-label pairs $(\mathbf{x}_1, y_1) \ldots (\mathbf{x}_n, y_n)$. And the model is linear regression: $y_i = \mathbf{w}^T\mathbf{x}_i$. ϵ is the allowed error margin (a constant), and $\xi_{i,+}, \xi_{i,-}$ are the slack variables.

1. Show that the constrained optimization problem in Equation (1) is equivalent to the following unconstrained optimization problem.

$$\text{minimize } \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\left[\underbrace{\max(0, |\mathbf{w}^T\mathbf{x}_i - y_i| - \epsilon)}_{\ell(\mathbf{w};\mathbf{x}_i,y_i)}\right] \tag{2}$$

5 pts.

2. Compute the gradient of the loss function (where it is differentiable).

5 pts.

3. Write the pseudo code for using stochastic gradient descent to derive the optimal parameters.

7 pts.

Question 3: Kernels (17 pts.)

1. Determine which of the following functions are kernel functions and which are not. Provide formal arguments to support your answer. Assume throughout that $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ with $\mathbf{x} = (x_1, \ldots, x_d)$ and $\|\cdot\|$ to be the Euclidean norm.

3+2+2+3 pts.

(a) $k(\mathbf{x}, \mathbf{y}) = \sqrt{\|\mathbf{x} - \mathbf{y}\|^2 + 1}$.

(b) $k(\mathbf{x}, \mathbf{y}) = C + \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|}$ for any constant $C \geq 0$.

(c) $k(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{d} (x_i + y_i)$.

2. In this problem we will look at kernel functions defined over finite sets. Let $\mathcal{X} = \{1, 2, \ldots, n\}$ so that $2^{\mathcal{X}}$ denotes the set of all subsets of $\mathcal{X}$. Show that $k : 2^{\mathcal{X}} \times 2^{\mathcal{X}} \to \mathbb{R}$ where

$$k(A, B) = \exp(-\frac{1}{2}|A \triangle B|), \quad \forall A, B \in 2^{\mathcal{X}}$$

is a kernel function. Here $A \triangle B = (A \cup B) \setminus (A \cap B)$ is the symmetric difference of A and B.

Hint: If $k(x, y)$ is a kernel then $\exp(k(x, y))$ is also a kernel.

7 pts.

Question 4: Probabilistic Modelling (17 pts.)

1. Apply Bayesian decision theory to derive the optimal action for some input $\mathbf{x}'$ in the context of logistic regression with asymmetric cost.

   - The estimated conditional distribution is given by

$$\hat{P}(y \mid \mathbf{x}) = \begin{cases} \sigma(\mathbf{w}^T\mathbf{x}) & \text{if } y = 1 \\ 1 - \sigma(\mathbf{w}^T\mathbf{x}) & \text{if } y = -1 \end{cases}$$

   - Action set $\mathcal{A}\{+1, -1\}$
   - Cost function given by

$$cost(y, a) = \begin{cases} a & \text{if } y = 1, a = -1 \\ b & \text{if } y = -1, a = 1 \\ 0 & \text{otherwise.} \end{cases}$$

   Given $\mathbf{w}$ and a new point $\mathbf{x}$, which action is chosen by Bayesian decision theory for the above asymmetric cost function?

8 pts.

2. Suppose that you are a part of a team that has trained n weather prediction models. The models use readings from a set of sensors that measure weather conditions on a given day and predict the temperature for the following day. Each model is fully determined by a vector of parameters $\mathbf{w}_i$ and estimates the conditional probability $P(y \mid \mathbf{x}, \mathbf{w}_i)$ of observing temperature y if the sensor state is $\mathbf{x}$. Furthermore, based on historical data your team has a prior belief over the models given by $P(\mathbf{w}_i) = \frac{2i}{n(n+1)}$.
Given that the measured sensor state today is $\mathbf{x}'$ please calculate the predictive distribution $P(y' \mid \mathbf{x}')$ for the temperature tomorrow using:

a) MAP estimation

3 pts.

b) Bayesian Model Averaging

5 pts.

Question 5: Gaussian Processes (17 pts.)

1. Let $P(f) \sim GP(f; 0, k)$. We observe $y = f(x) + \epsilon$, where $x, y \in \mathbb{R}$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

   (a) Assume a polynomial kernel $k(x, x') = (xx' + 1)^d$, $d = 2$.

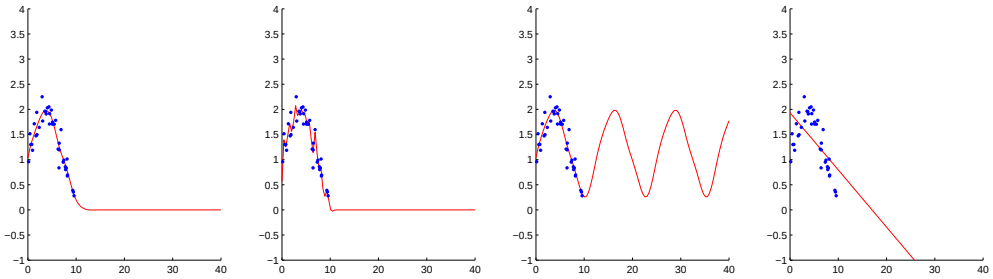
      (a) Compute $P(f(x')|x, y) = \mathcal{N}(\mu_{x'|x}, \sigma^2_{x'|x})$.

      (b) Compute $\lim_{\sigma^2 \to 0} \mu_{x'|x}$ and $\lim_{\sigma^2 \to 0} \sigma^2_{x'|x}$.

   (b) Now consider what happens to the posterior mean $\mu_{x'|x}$ obtained in 1(b) and as $x' \to \infty$ for some fixed $x \in \mathbb{R}$. In particular, compute the limit for $\mu_{x'|x}/x'$. Interpret the results.

7 pts.

2. At the plot below you can see four different predictions made using Gaussian Processes regression based on the same set of data but using different kernels under the assumption that noise parameter $\sigma = 0.1$



(a) Specify which kernel was used for each of these four estimations. You can indicate your choice by writing the number of the kernel next to the corresponding plot:

(1) $k(x, x') = \exp(-(x - x')^2)$

(2) $k(x, x') = xx'$

(3) $k(x, x') = (xx' + 1)^d$, $d = 3$

(4) $k(x, x') = \exp(-\sin^2(x - x'))$

(5) $k(x, x') = \exp(\frac{-(x-x')^2}{4.48})$

(6) $k(x, x') = xx' + \exp(-(x - x')^2)$

(7) $k(x, x') = xx' + 1$

(b) Draw (approximately) the variance of each estimate given the chosen kernel.

10 pts.

Question 6: Boosting (17 pts.)

Consider the AdaBoost algorithm shown below to answers questions 1 to 3 of this section.

Input: Data set $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$

Initialize $w_1^{(1)} = \cdots = w_n^{(1)} = 1$

For $i = 1 : m$

Train weak learner on weighted data: $h_i \leftarrow \arg\min_h \sum_{j=1}^{n} w_j^{(i)} [h(\mathbf{x}_j) \neq y_j]$

Compute $\text{err}_i = \dfrac{\sum_{j=1}^{n} w_j^{(i)} [y_j \neq h_i(\mathbf{x}_j)]}{\sum_{j=1}^{n} w_j^{(i)}} \quad \beta_i = \log \dfrac{1 - \text{err}_i}{\text{err}_i}$

Update all weights:

$$w_j^{(i+1)} = w_j^{(i)} \exp(\beta_i [h_i(\mathbf{x}_j) \neq y_j])$$

Output: $f(\mathbf{x}) = \sum_{i=1}^{m} \beta_i h_i(\mathbf{x})$

1

1. Explain in what sense AdaBoost is a greedy algorithm and why.

2 pts.

2. List the parameters that need to be manually tuned in AdaBoost.

2 pts.

3. What does AdaBoost do if $\text{err}_i = 0.5$? Explain why.

2 pts.

4. The AdaBoost algorithm given above fails if $\text{err}_i = 0$. How can you fix this? What should the algorithm do in this case?

2 pts.

5. Mark the following statements about AdaBoost with T if they are true or with F if they are false.

   ☐ Adaboost trains classifiers using only the points that were misclassified in the previous iteration.

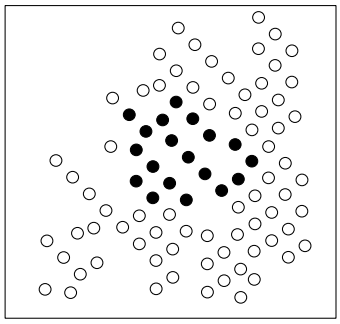
   ☐ If training error is zero, test error is guaranteed to be zero as well.

   ☐ Training error always goes down when more weak classifiers are added.

   ☐ At iteration $i = p$, err_p indicates the probability that a new point $\mathbf{x}$ will be misclassified by $\text{sgn}(f(\mathbf{x}))$, where $f(\mathbf{x}) = \sum_{i=1}^{p} \beta_i h_i(\mathbf{x})$.

4 pts.

6. Consider a binary classification problem in a two dimensional feature space. For the training points shown in the figure select the $minimum$ number of classifiers that are needed for a boosting algorithm to achieve zero training error with the indicated type of classifier. If the type of classifier can never achieve zero training error, select the last option.



(a) Logistic regression

- ☐ 1
- ☐ 2 to 5
- ☐ more than 5
- ☐ Not possible

(b) SVM with Gaussian kernel

- ☐ 1
- ☐ 2 to 5
- ☐ more than 5
- ☐ Not possible

(c) Decision trees

- ☐ 1
- ☐ 2 to 5
- ☐ more than 5
- ☐ Not possible

(d) Decision stumps

- ☐ 1
- ☐ 2 to 5
- ☐ more than 5
- ☐ Not possible

1+1+1+1 pts.

Question 7: K-Means and EM (17 pts.)

1. Mark the following statements about clustering with T if they are true or with F if they are false.
   - ☐ The EM algorithm can only be used to fit a Gaussian Mixture Model.
   - ☐ K-means results can change after normalizing the data to zero mean, unit variance.
   - ☐ The EM algorithm for GMM does not depend on initialization.
   - ☐ Every step of the K-means algorithm is guaranteed not to increase the value of the objective function.
   - ☐ The EM-algorithm finds the global maximum of the likelihood.

   3 pts.

2. Given a dataset of five points in $\mathbb{R}$: $X = (1, 5, 7, 8, 8)$, and the number of clusters $K = 2$, run the K-means clustering algorithm by hand. Start with centers $c_1 = 1$, $c_2 = 10$. What are the final centers and the assignement of points to clusters?

   2 pts.

3. You are given a dataset of points in $\mathbb{R}^2$ and you want to find K clusters. Assume that you try both clustering with K-means and fitting a GMM with K components on the same dataset. Assume that all the cluster centers μ ended up in exactly the same points for both K-means and GMM. Is it possible that the decision boundaries between the clusters are different in these two cases?

Since GMM provides probabilistic assignment, consider that for GMM you assign a point $\mathbf{x}$ to cluster k if the probability of $\mathbf{x}$ being generated from the k-th component is the largest $\forall k = 1, \ldots, K$. Provide a justification for your answer.

3 pts.