ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Introduction to Machine Learning
2010

Prof. J. M. Buhmann

Final Exam

January 29th, 2011

First and Last name: ______________________________

ETH number: ______________________________

Signature: ______________________________

	Topic	Max. Points	Points Achieved	Visum
1	Bayesian Inference	25		
2	Regression	25		
3	Bagging and Boosting	25		
4	SVM	25		
5	Mixture Models	25		
Total		125		

Grade: ....

This page has been intentionally left blank.

Question 1: Bayesian Inference (25 pts.)

Consider the task of estimating mean μ and variance σ^2 of a Gaussian density from n i.i.d. observations $\mathcal{X} = \{x_1, \dots, x_n\}$, $x \in \mathbb{R}$.

a) Write the maximum likelihood estimator for the mean $\widehat{\mu}_{\text{ML}}(\mathcal{X})$ as an explicit function of the i.i.d. observations $\mathcal{X} = \{x_1, \dots, x_n\}$:
(please write the direct closed-form solution)

1. $\widehat{\mu}_{\text{ML}}(\mathcal{X}) = \arg\max p(\mathcal{X}|\mu) = \frac{1}{n} \sum_{i}^{n} x_i$ ✓ **1 pts.**

2. Is this a biased estimator? YES\NO ✓ **1 pts.**

b) Write the maximum likelihood estimator for the variance $\widehat{\sigma}^2_{\text{ML}}(\mathcal{X})$ as an explicit function of the i.i.d. observations $\mathcal{X} = \{x_1, \dots, x_n\}$:
(please write the direct closed-form solution) ✓

1. $\widehat{\sigma}^2_{\text{ML}}(\mathcal{X}) = \arg\max p(\mathcal{X}|\sigma^2) = \frac{1}{n} \sum_{i}^{n} (x_i - \mu)^2$ **1 pts.**

unbiased estimate of population variance: $\frac{1}{n-1} \sum_{i}^{n} (x_i - \mu)^2$

2. Is this a biased estimator? ✓YES\NO ✓ **1 pts.**

Now, assume that the variance σ^2 is known.

c) Given a Gaussian prior over the mean (prior with zero mean and variance one), write the posterior density $p(\mu|\mathcal{X})$ as an explicit function of the single observation $\mathcal{X} = \{x_1\}$:
(please write the direct closed-form solution)

$P(\mu) = \frac{1}{\sqrt{2\pi}} \exp \frac{-(x_i)^2}{2}$ $\mu = 0$, $\sigma^2 = 1$

bc conjugate pair; write posterior distribution $p) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x_i - \mu)^2}{2\sigma^2}$

1. $p(\mu|\mathcal{X}) = \frac{p(\mathcal{X}|\mu)p(\mu)}{p(\mathcal{X})} = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x_i)^2}{2} - \frac{(x_i - \mu)^2}{2\sigma^2}$ **2 pts.**

$= \boxed{\frac{1}{(2\pi)^{\frac{1}{4}}\sigma} \exp \frac{-(x_i^2\sigma^2 + (x_i - \mu)^2)}{2\sigma^2}}$ $\quad -\frac{x_i^2}{2} - \frac{(x_i - \mu)^2}{2\sigma^2} = \frac{-x_i^2\sigma^2 - (x_i - \mu)^2}{2\sigma^2}$

2. Is the posterior a Gaussian density? YES\NO ✓ **1 pts.**

3. Is the posterior a Gaussian density for all priors? YES\NO **1 pts.**

you have to consider all existing priors, not only gaussians

d) Given a Gaussian prior over the mean (prior with zero mean and variance one), write the maximum a posteriori estimator $\widehat{\mu}_{\text{MAP}}(\mathcal{X})$ as an explicit function of the i.i.d. observations $\mathcal{X} = \{x_1, \ldots, x_n\}$ (recall that $\int e^{-(\alpha^2)} d\alpha = \sqrt(\pi))$:
(please write the direct closed-form solution)

Not sure why you want to integrate instead of take derivative.

1. $\widehat{\mu}_{\text{MAP}}(\mathcal{X}) = \arg\max p(\mu|\mathcal{X}) = \frac{n}{n+\sigma^2}\hat{\mu} + \frac{\sigma^2}{n+\sigma^2}\mu \quad \sum_{i}^{n} x_i$ **2 pts.**

2. Does the maximum a posteriori estimate equal the mean of the posterior?
YES\NO $\hat{\mu}_{MAP} = \mu_n$ **1 pts.**

GMM

Now consider a binary classification task from observations $\mathcal{X} = \{x_1, \ldots, x_n\}$, with $x \in \mathbb{R}^D$. 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 $i = 1, 2$).

e) Write the discriminant $g_{y_1}(x) = p(y_1|x)$ as an explicit function of class prior, mean and covariance:
(please write the direct closed-form solution) $\sum_{i=1}^{2} \pi_i \, p(x|\mu_i, \Sigma_i)$

$$g_{y_1}(x) = \frac{P(x|y_1, \mu_1, \Sigma_1) \cdot P(y_1|\mu_1, \Sigma_1)}{P(x|\mu_1, \Sigma_1)}$$ **3 pts.**

$$= \frac{1}{(2\pi)^{d}|\Sigma_1|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right) \cdot \pi_1$$

f) Assume that $\Sigma_1 = \Sigma_2 = \sigma^2\mathbb{I}$, where $\mathbb{I}$ denotes the identity matrix. Write the equation satisfied by the separating decision surface. The equation must be an explicit function of x_1 (the single observation), of class prior, means, and covariance:
(please write the direct closed-form solution)

1. $0 = g_{y_1}(x) - g_{y_2}(x) = \left[\frac{1}{(2\pi)^{d/2}|\Sigma_1|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_1-\mu_1)^T\Sigma_1^{-1}(x_1-\mu_1)\right) \cdot \pi_1\right]$ **2 pts.**

$$- \left[\frac{1}{(2\pi)^{d/2}|\Sigma_2|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_1-\mu_2)^T\Sigma_2^{-1}(x_1-\mu_2)\right) \cdot \pi_2\right]$$

Every time $\Sigma_i = \Sigma_j$ then → linear.

Duda pg 36

2. In this case, is the decision surface constant, linear, quadratic, cubic, or something else? constant\linear\quadratic\cubic\other **2 pts.**

3. Would it be possible to obtain a cylindrical decision surface? YES\NO **1 pts.**

Be linear.

$-\frac{1}{2} x_1^T (\Sigma_1^{-1} - \Sigma_2^{-1}) x_1$

$a x_1 +$ ______

g) In the two dimensional case subject to uniform class prior, write means $(\bar{\mu}_1, \bar{\mu}_2)$ and covariances (Σ_1, Σ_2) such that the decision surface is a hyper-plane:
(any numerical instantiation which satisfies this constraint is acceptable)

$$\mu_1 = \frac{\sum_i x_i P(\mu_1^{old}, \Sigma_1^{old} | x_i)}{\sum_i P(\mu_1^{old}, \Sigma_1^{old} | x_i)} \qquad \frac{1}{N} \Sigma x_i \qquad \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\mu_2 = \frac{\sum_i x_i P(\mu_2^{old}, \Sigma_2^{old} | x_i)}{\sum_i P(\mu_2^{old}, \Sigma_2^{old} | x_i)} \qquad \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

2×1

$$\Sigma_1 = \Sigma_2 \frac{\sum_i (x_i - \mu_1^{new})^2 \cdot P(\mu^{old}, \Sigma_1^{old} | x_i)}{\sum_i P(\mu^{old}, \Sigma^{old} | x_i)}$$

equal

$$\Sigma_2 = \Sigma_1 \frac{\sum_i (x_i - \mu_2^{new})^2 \cdot P(\mu^{old}, \Sigma_2^{old} | x_i)}{\sum_i P(\mu^{old}, \Sigma^{old} | x_i)}$$

2×2

3 pts.

h) In the two dimensional case subject to uniform class prior, write means (μ_1, μ_2) and covariances (Σ_1, Σ_2) such that the decision surface is spherical:
(any numerical instantiation which satisfies this constraint is acceptable)

pg. 41.

$\mu_1 = \begin{bmatrix} \quad \end{bmatrix}$

$\mu_2 = \begin{bmatrix} \quad \end{bmatrix}$

$\Sigma_1 = \begin{bmatrix} \quad \end{bmatrix}$

$\Sigma_2 = \begin{bmatrix} \quad \end{bmatrix}$

3 pts.

This page has been intentionally left blank.

Question 2: Regression (25 pts.)

Consider the linear regression model expressed in the homogeneous coordinates:

$$y = \boldsymbol{\beta}_{(0)} + \sum_{i=1}^{D} \mathbf{x}_{(i)}\boldsymbol{\beta}_{(i)} = \mathbf{x}^\top \boldsymbol{\beta},$$

where $\mathbf{x} = (1, \mathbf{x}_{(1)}, ..., \mathbf{x}_{(D)}) \in \mathbb{R}^{D+1}$ is the input variable and y is the corresponding target variable.

Assume that the input dataset is given by the matrix $\mathbf{X} \in \mathbb{R}^{N\times(D+1)}$ whose first column is $\mathbf{1}$. Then the linear regression model for all the observations is written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$. Consider this linear regression model and answer the following questions:

a) For this problem, formally define the Residual Sum of Squares (RSS) cost function and write it down in matrix notation.

$RSS = (y_{True} - \hat{y})^2$

Answer:

$$RSS = (y_{True} - \beta^T x)^T (y_{True} - \beta^T x)$$

2 pts.

b) Briefly motivate the connection between minimizing the Residual Sum of Squares and maximizing the likelihood of $\mathbf{y}$ given $\mathbf{X}$.

Answer:

To minimize RSS is to reduce the error between the regressor and data points

maximizing likelihood $\max_{\beta} P(X|\beta)$ ⇒ finding the Betas coefficients of the regressor that will "best fit" the data.

Error of probability $P(\varepsilon_1, \dots, \varepsilon_n | \beta) = \prod_{i=1}^{n} P(\varepsilon_i|\beta) = \prod_{i=1}^{n} P(y_i - x^T\beta|\beta)$

2 pts.

$Y = X\beta + \varepsilon$ $\quad \varepsilon_i \sim N(0, 6^2)$

$Y \sim N(XB, 6^2)$ $\quad \frac{1}{\sqrt{2\pi 6^2}} \exp\left(\frac{-(Y_i\ XB)}{26^2}\right)$

$Y\ [Y_1 \dots Y_N]$

c) We minimize the RSS function and infer the model parameters as $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. Considering the matrix inverse operation, mathematically describe why in practice we are interested in regularized models such as *ridge* regression models rather than the given unregularized model.

$\hat{\beta} = (X^TX + \lambda I)^{-1} X^T Y$

we can directly manipulate size of β.

Good for stabilizing matrix

Positive-definite.

Full rank.

no corre between samples

Answer: $RSS(\beta) = (Y - \beta^T X)^T (Y - \beta^T X)$ $\qquad \|\beta\|_R = \left(\sum_i |\beta_i|^2\right)^{\frac{1}{2}}$

In regularization, you assume β to be distributed by nature according to gaussian law.

to choose β values that control complexity of model.

β is distributed by nature, $p(\beta) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\sigma_\beta^2}} \cdot \exp\left(\frac{-\beta^2}{2\sigma_\beta^2}\right) \Rightarrow$ Take log of $P(\beta)$, $L_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$

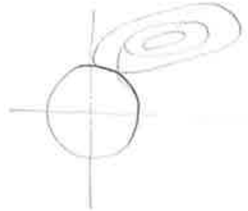
then arrive at $-\log p(\beta) = \sum_{i=1}^{d} \frac{\beta^2}{2\sigma_\beta^2} + \text{constant} \Rightarrow \underset{\beta}{\arg\min} \left\{ RSS(\beta) + \lambda \sum_{j=1}^{d} \beta^2 \right\}$

λ = larange multiplier

3 pts.

d) Formally define ridge and LASSO regression models. By depicting appropriate plots, demonstrate the difference between these two models in inferring the model parameters $\hat{\boldsymbol{\beta}}$.

Answer:

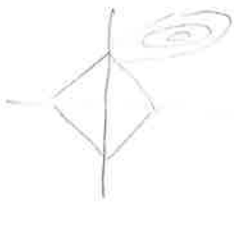


Ridge regression

X

L2-penalty $\quad \|\hat{\beta}\|_2 = \left(\sum_i |\beta_i|^2\right)^{\frac{1}{2}}$

shrinks the regression coefficients by imposing a penalty on their size.



LASSO → L1 penalty $\quad \|\hat{\beta}\|_1 = \left(\sum_i |\beta_i|\right)$

favors β values to be exactly ϕ when λ is large

Sparseness penalty model with few coefficients non-vanishing.

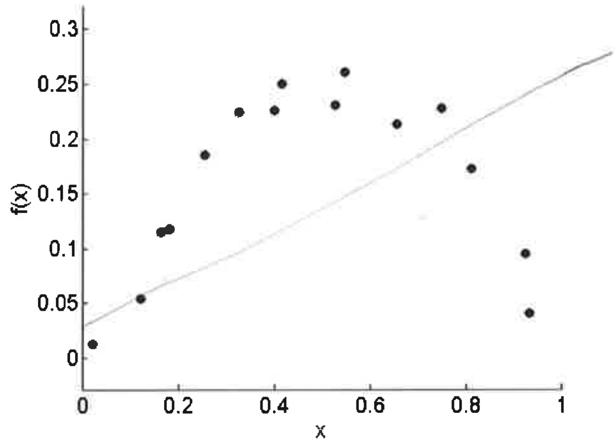
Because the least-square surface often hits the corners of the constraint surface.

$\varphi_1 \beta_1 + \varphi_2 \beta_2 = \varphi$

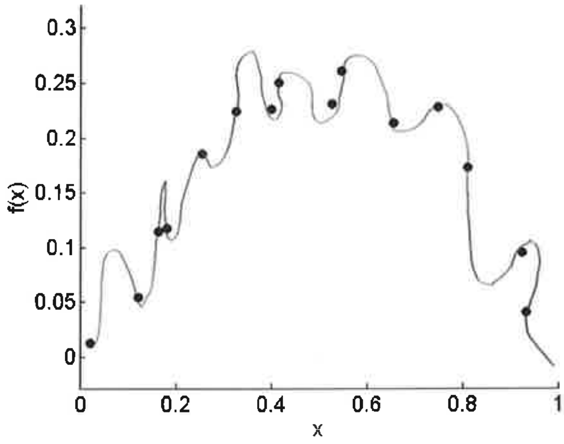
5 pts.

Now, we consider a general form of the problem where the set of observation $\{(\mathbf{x}_i, y_i)\}_{i=1}^n, \mathbf{x}_i \in \mathbb{R}^D, y_i \in \mathbb{R}$ are drawn i.i.d. from the joint distribution $P(X,Y)$. The goal is to find the regression function $f \in \mathcal{F}$ such that the mean squared error $\mathbb{E}_{XY}[(Y - f(X))^2]$ is minimal. Where the hypothesis class $\mathcal{F}$ contains the set of all polynomial functions.

e) Choosing an inappropriate function f might lead to either underfitting or overfitting. For the depicted given data, draw relevant plots to show each of these situations. For the overfitting case, briefly explain what happens if we have more observations.
Answer:



underfitting



overfitting

4 pts.

We can split the mean squared error $\mathbb{E}_{XY}[(Y - f(X))^2]$ into bias and variance, and find the best tradeoff between them.

f) Write down the mathematical definition of bias and variance.

Answer: MSE = Bias^2 + variance.

$$MSE = \left(\mathbb{E}[\hat{f}(x)] - f(x)\right)^2 + \mathbb{E}\left[\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\right]^2$$

Bias $= \mathbb{E}[\hat{f}(x)] - \underbrace{f(x)}_{Y}$ variance $= \mathbb{E}\left[\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\right]^2$

2 pts.

g) Expand the mean squared error $\mathbb{E}_{XY}[(Y - f(X))^2]$ and write it in the form of variance + squared bias.

Answer: ϕ $\hat{\phi}$

$$\mathbb{E}_{xy}[(Y - f(x))^2] = \mathbb{E}_{x,y}[f(x)^2 - 2Yf(x) + Y^2]$$

$$= \mathbb{E}_{x,y}[f(x)^2]$$

3 pts.

In this section we study how the averaging changes the bias of a set of unbiased estimators. Assume that we are given a set of B estimators $\hat{f}_1(\mathbf{x}), \hat{f}_2(\mathbf{x}), ..., \hat{f}_B(\mathbf{x})$. We take the average estimator by $\hat{f}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(\mathbf{x})$.

h) Calculate the bias of the average estimator $\hat{f}(\mathbf{x})$ in terms of the bias of the estimators $\hat{f}_1(\mathbf{x}), ..., \hat{f}_B(\mathbf{x})$.
Hint: Start with the mathematical definition of the bias for $\hat{f}(\mathbf{x})$.
Answer:

$$\text{bias}(\hat{f}(x)) = \mathbb{E}[\hat{f}(x)] - \mathbb{E}[Y|X] = \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(x)\right] - \mathbb{E}[Y|X]$$

$$= \frac{1}{B}\sum_{b=1}^{B}\mathbb{E}[\hat{f}_b(x)] - \mathbb{E}[y|X] = \boxed{\frac{1}{B}\sum_{b=1}^{B}\text{bias}(\hat{f}_b(x))}$$

✓

3 pts.

i) According to the results, briefly explain why unbiased estimators remain unbiased after averaging.
Answer:

✓

Because taking the average of something gives you an unbiased. And in this case, if the estimator is already unbiased and taking the average of the different unbiased estimators again is unbiased.

1 pts.

This page has been intentionally left blank.

Question 3: Bagging and Boosting (25 pts.)

Bagging and *boosting* are two possible approaches to combine multiple models (classifiers or regressors) to achieve a *composite model* with improved performance. Both methods can be used in both a classification and a regression setting.

a) State two essential differences between bagging and boosting.

Answer:

	Bagging	Boosting
1)	Bagging varies the training sets using resampling	Trains weak learners with the same training set on every iteration
2)	gives the same importance to every prediction	weights the prediction of every weak classifier according to its accuracy,

2 pts.

b) Explain in terms of the bias-variance trade-off why the idea of combining models works.

Answer:

without combining models, either you are trying to battle with having a high bias and low variance or a low bias and high variance.

when you combine models, you are gaining a diverse group of classifiers that partition the sample space in a complex manner that tends to reduce boths the bias and variance.

Bagging → reduce variance

Boosting reduce bias.

Give more weight to misclassified.

2 pts.

c) We now look at the problem of regression and how the combination of individual regression models can give better results. Left-column figures show the individual regression models, right-column figures show the true target function (solid) and the output of averaging the individual models to obtain a composite model (dashed).

1. The individual regression models (in Figure 1 and Figure 3) have been regularized using a regularization parameter λ, i.e. the cost function had the following form

$$RSS_{Ridge}(\boldsymbol{\beta}) = \sum_{i=1}^{n}(t_i - \phi(\mathbf{x}_i)^\top \boldsymbol{\beta})^2 \ + \ \lambda||\boldsymbol{\beta}||_2^2$$

   The parameters used were $\lambda = 0.09$ and $\lambda = 13.5$.
   Associate the regularization parameters to the figures in the left column.
   Answer:

   large λ pulls the weight parameters toward

   Figure 1: $\lambda = \ldots$ 13.5 → zero leading to large bias.

   Figure 3: $\lambda = \ldots$ 0.09 → Allows model become finely tuned to noise → leading to large variance

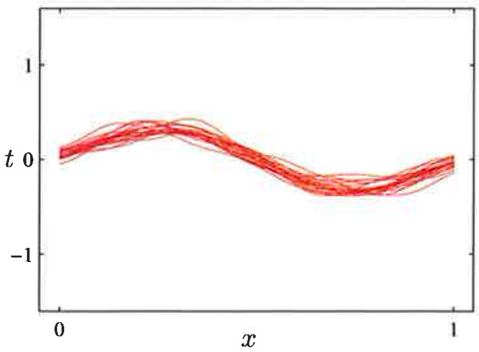
2. Please interpret the figures in terms of the bias-variance trade-off
   Answer:

   Figure 1 and 2: Bias is high, variance is low
   rigid model.

   Figure 3
   Figure 3 and 4: Bias is low, variance is high

   Figure 4 gives a low bias and variance.
   flexible models

Figure 1



⇒

Figure 2

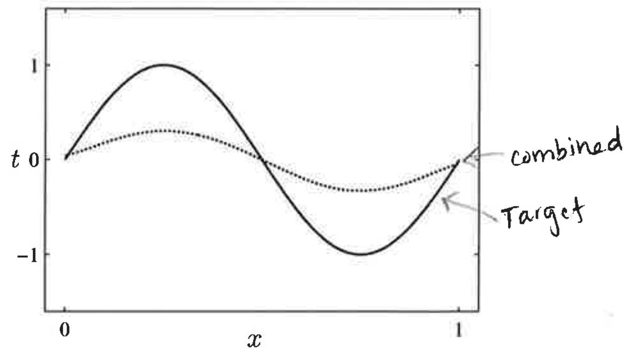
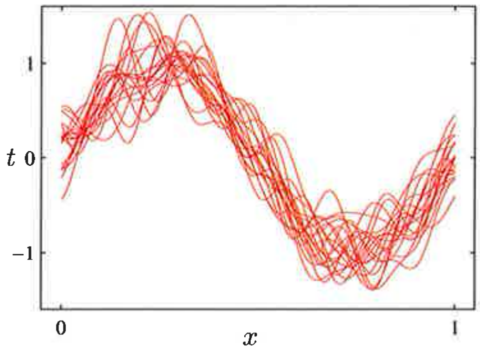
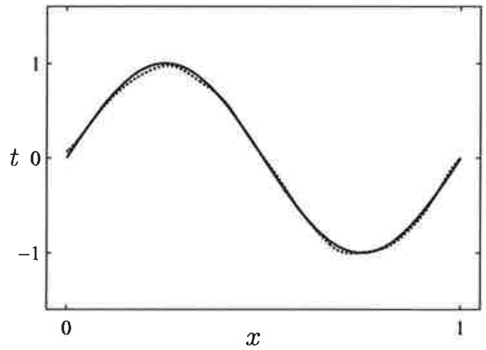


Figure 3



⇒

Figure 4



4 pts.

d) We now consider *bagging* in a regression setting. In order to model a *target function* $h(x)$, we combine M *individual models* $y_m(x), m = 1..M$ to obtain a *committee model* $y_{COM}(x)$. We model the error of each individual model using $\epsilon_m(x), m = 1..M$. This question's final goal is to show that the error of the committee model is M times smaller than the average error of the individual models.

1. Write down the output of the committee model, which averages the individual model's outputs. Answer:
   $y_{COM}(x) = \ldots \frac{1}{M}\sum_{m=1}^{M} y_m(x)$ ✓
2. Write down the error $\epsilon_m(x)$ of an individual model in terms of the target function $h(x)$ and the output of the individual model $y_m(x)$. Answer:
   $\epsilon_m(x) = h(x) - y_m(x)$ ✓
3. Write down the expected squared error of an individual model. $\mathbb{E}_x$ denotes the expectation value w.r.t. the distribution of x. Answer:
   $E_m = \mathbb{E}_x\left[\varepsilon_m(x)^2 \right]$ ✓
4. Write down the average of the expected squared errors made by the individual models ✓
   Answer:
   $E_{AV} = \ldots \frac{1}{M}\sum_{m=1}^{M} E_x\left[\varepsilon_m(x)^2\right]$
5. Write down the expected squared error made by the committee model, in terms of the output of the committee model and the target function. ✓
   Answer:
   $E_{COM} = \mathbb{E}_x\left[\left(h(x) - y_{com}(x) \right)^2\right]$ $\quad E_x\left[\left(\frac{1}{M}\sum_m^M y_m(x) - h(x)\right)^2\right]$
6. Show that $E_{COM} = \frac{1}{M}E_{AV}$.
   Hint: Use the assumption that the errors of the individual models are uncorrelated, i.e. $\mathbb{E}_x[\epsilon_m(x)\epsilon_l(x)] = 0, m \neq l$, and that the mean error of the individual models is zero, i.e. $\mathbb{E}_x[\epsilon_m(x)] = 0$.
   Answer:
   $E_{COM} = E_x\left[\left(\frac{1}{M}\sum_m^M y_m(x) - h(x)\right)^2\right]$ $\quad \left(\frac{1}{M}\sum_m y_m(x) - h(t)\right)$
   $E_{com} = \left(\frac{1}{M}\sum_m E_x\left[y_m(x) - h(x)\right]\right)^2$ $\quad \left(\frac{1}{M}\sum y_m\right.$
   $= \left(\frac{1}{M}\sum_m E_x\left[\epsilon_m(x)\right]\right)^2$

10 pts.

$E_{COM} = \frac{1}{M}\sum_{m} E_{AV}$ $\quad E\left[y_{com}\right] = \frac{1}{M^2}\sum_m E\left[e_i(x)^2\right]$

$\sum_i\sum_j E(e_i(x)\, e_j(x).)$

$E\left[\left(\frac{1}{M}\sum_m y_m(x) - h(x)\right)\left(\frac{1}{M}\sum_m y_m(x) - h(x)\right)\right]$

$a^2 + 2ab + b^2$

y_i^2

$\frac{1}{M^2}E\left[\ \right]$

We now turn our attention to *boosting* in a classification setting.

e) Given i.i.d. training data $\{(\mathbf{x}_1, c_1), (\mathbf{x}_2, c_2), ..., (\mathbf{x}_n, c_n)\}$, $\mathbf{x}_i \in \mathbb{R}^D$, $c_i \in \{-1, 1\}$, and individual classifiers $y_m(\mathbf{x}), m = 1...M$, we can define the following additive model

$$f_k(\mathbf{x}) = \frac{1}{2}\sum_{j=1}^{k} \alpha_j y_j(\mathbf{x})$$

where α_j are weights, $k \leq M$.
Think of $f_k(\mathbf{x})$ as the *additive model* which uses the first $1..k$ individual classifiers.

1. Define the sum-of-squares error of $f_k(\mathbf{x})$ on the training dataset.
Answer:
Here we consider an exponential error fxn

$Err_k = \quad ... \quad \cancel{\sum_{i=1}^{n} \exp\left(-c_i f_k(x_i)\right)}$

2. Assume that the first $1..(k-1)$ classifiers $y_j(\mathbf{x})$ and weights α_j, $j = 1..(k-1)$ have already been determined, i.e. fixed.

Show that finding the optimal k-th classifier $y_k(\mathbf{x})$ and its weight α_k involves fitting the k-th classifier to the residual errors $f_{k-1}(\mathbf{x}_i) - c_i$ made by the $(k-1)$-th additive model.

$y_k(x) = ? \quad \alpha_k = ?$

(*Hint: Start from the expression Err_k and recognize the terms $f_{k-1}(\mathbf{x}_i) - c_i$*)

$\sum_{i=1}^{n} \exp(-c_i f_k(x_i))$

$\alpha = (f_{k-1}(x_i) - c_i)$

$y_k(x$

$\boxed{\sum_{i}^{n} (c_i - f_k(x_i))^2}$

$\sum_i \left(f_{k-1}(x_i - c_i) - f_k(x) \right)^2$

7 pts.

This page has been intentionally left blank.

Question 4: SVM (25 pts.)

In the following questions choose one answer only.

[1, .. d]

a) Why does a Support Vector Machine generalizes better than a Perceptron?

1. The selected support vectors tend to be very typical samples. border case.
2. By supporting vectors it is not restricted to scalar input.
3. The requirement of maximal margins reduces the arbitrariness.
4. The support vector machine will have access to more training data.

2 pts.

b) What is the advantage of using kernels in support vector machines?

1. They tend to maximize the margins.
2. They enable non-linear separations.
3. They will increase the number of support vectors.
4. They reduce the risk of getting stuck in local minima.

2 pts.

Explain your answer to the following questions in 1-2 sentences.

c) Suppose that you have a binary SVM classifier with a linear kernel. Consider a vector $\mathbf{x}_i$ for which $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0) > 1 \quad (\mathbf{x}_i \in \mathbb{R}^D, y_i \in \{-1, +1\})$

1. Is $\mathbf{x}_i$ correctly classified? YES\NO
2. If we remove $\mathbf{x}_i$ from the training set and re-train the classifier, will the decision boundary change or stay the same?

Answer:

The decision boundary might change depending if the training sample was a support vector in the previous case.

3 pts.

d) We would like to build a binary SVM classifier for histograms. Let H denote the class of histograms, for each $\mathbf{h} \in H$ the following properties hold:

class of histograms

1. $\mathbf{h} = \{h_1, \ldots, h_m\}, \quad \forall j : h_j \geq 0$
2. $\sum_{j=1}^{m} h_j = L$ for some $L \in \mathbb{N}$.

Consider using the following kernel function:

$$K(\mathbf{h}, \tilde{\mathbf{h}}) = \sum_{j=1}^{m} \min(h_j, \tilde{h}_j) \quad \mathbf{h}, \tilde{\mathbf{h}} \in H$$

?

What kind of separation is imposed by this kernel? (which points in the space will become close/far using this kernel).

Answer:

3 pts.

e) Let $\mathcal{X}$ be a finite set and consider the following function. For $\mathbf{x}, \bar{\mathbf{x}} \in \mathcal{X}$:

$$K(\mathbf{x}, \bar{\mathbf{x}}) = \begin{cases} 1 & \text{if } \mathbf{x} = \bar{\mathbf{x}} \\ 0 & \text{else} \end{cases}$$

Prove that $K(\mathbf{x}, \bar{\mathbf{x}})$ is a legitimate kernel function.

Hint: One possible way to prove it is to find a feature mapping $\phi(\mathbf{x})$, such that

$K(\mathbf{x}, \bar{\mathbf{x}}) = \langle \phi(\mathbf{x}), \phi(\bar{\mathbf{x}}) \rangle$.

$x = (x_1, x_2)$

$\bar{x} = (\bar{x}_1, \bar{x}_2)$

Answer:

$\langle \phi(x), \phi(\bar{x}) \rangle = \phi(x_1, x_2) \cdot \phi(\bar{x}_1, \bar{x}_2)$

3 pts.

$L2$-SVM use the square sum of the slack variables ξ_i in the objective function instead of the linear sum of the slack variables (squaring the hinge loss). Let $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be a training set of examples and binary labels $y_i \in \{-1, +1\}$. The primal formulation of the $L2$-SVM is as follows

$$\min_{\mathbf{w}, w_0, \xi} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{n} \xi_i^2$$

$$w = \sum_{1}^{n} \alpha_i$$

s.t.

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + w_0) \geqslant 1 - \xi_i \quad i = 1, \ldots, n$$
$$\xi_i \geqslant 0 \quad i = 1, \ldots, n$$

$$y_i(\langle w, x_i \rangle + w_0) - 1 + \xi_i$$

f) Show that removing the last set of constraints: $\forall i : \xi_i \geqslant 0$ does not change the optimal solution to the primal problem.

Answer:

$$\lambda y_i w^T x + w_0 - 1$$

$$L(w, w_0, \xi, \lambda_1, \lambda_2) = \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_i^n \xi_i^2 + \lambda_1(y_i(\langle w, x_i \rangle + w_0) - 1 + \xi_i) + \lambda_2(\xi) \stackrel{!}{=} 0$$

$$\frac{d}{d\lambda_1} = y_i(\langle w, x_i \rangle + w_0 - 1 + \xi_i) = 0$$

$$\frac{d}{d\lambda_2} = \xi = 0$$

$$\frac{d}{d\bar{w}} = \bar{w} + \lambda_1 y_i x = 0 \quad \lambda_1 = -\frac{\bar{w}}{y_i x}$$

$$\frac{d}{dw_0} = \lambda_1 = 0$$

$$= \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_i^n \xi_i^2 - \frac{\vec{w}}{y_i x}(y_i(w^T x_i + w_0) - 1 + \xi_i) + \cancel{\lambda_2(0)}.$$

4 pts.

w/out $\xi \geq 0$ constraint

$$= \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_i^n \xi_i^2 - \frac{\vec{w}}{y_i x}(y_i(w^T x_i + w_0) - 1 + \xi_i \; x)$$

Now we would like to derive the dual form of the $L2$-SVM.

g) Write down the Lagrangian of the $L2$-SVM.
Answer:

$$L(\qquad\qquad) =$$

2 pts.

h) Compute the derivatives of the Lagrangian with respect to the appropriate variables.
Answer:

3 pts.

We decided not to bother you with the remaining computation, instead we finished the derivation ourselves.

i) Below are 4 optimization problems, only 1 of which is the dual $L2$-SVM. Circle the optimization problem corresponding to the dual $L2$-SVM.
Explain your choice by either shortly falsifying the other options or by showing the full derivation.
Hint: The full derivation is more time consuming.

Derive dual form.

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j \mathbf{x}_i{}^T \mathbf{x}_j - \frac{1}{2C}\sum_{i=1}^{n}\alpha_i^2$$

s.t.:

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$
$$\forall i: \ \alpha_i \geq 0$$

(a)

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j \mathbf{x}_i{}^T \mathbf{x}_j$$

s.t.:

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \checkmark$$
$$\forall i: \ 0 \leq \alpha_i \leq \frac{\xi_i}{C}$$

(b)

no y's

$$\min_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j \mathbf{x}_i{}^T \mathbf{x}_j - \frac{1}{2C}\sum_{i=1}^{n}\alpha_i^2$$

s.t.:

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$
$$\forall i: \ \alpha_i \geq 0$$

(c)

$$\max_{\alpha,\xi} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j \mathbf{x}_i{}^T \mathbf{x}_j + C\sum_{i=1}^{n}\xi_i$$

slack

s.t.:

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$
$$\forall i: \ \alpha_i \geq 0$$

Doesn't include constraint for C.

(d)

Answer:

IN the primal form you solved for w's and pluged it into dual form $w = \sum_i \alpha_i y_i x_i$.

3 pts.

This page has been intentionally left blank.

Question 5: Mixture of Bernoulli models (25 pts.)

Let $\mathbf{x} = (x_1, \ldots, x_D)^T \in \{0,1\}^D$ be a D-dimensional random binary vector. We assume that every x_i is Bernoulli distributed with parameter μ_i, i.e.

$$p(x_i = 1;\, \mu_i) = \mu_i$$
$$p(x_i = 0;\, \mu_i) = 1 - \mu_i,$$

which can be also written as

$$p(x_i;\, \mu_i) = \mu_i^{x_i}(1-\mu_i)^{1-x_i}.$$

Under the assumption of the x_i's being independent, the distribution of the random vector $\mathbf{x} = (x_1, \ldots, x_D)^T$ is

$$p(\mathbf{x};\, \boldsymbol{\mu}) = \prod_{i=1}^{D} \mu_i^{x_i}(1-\mu_i)^{1-x_i},$$

$COV(X) = E[XX^T] - E[X]E[X]^T$

with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_D)^T$.

$E[\mu_i^{x_i}$

a) What is $\mathbb{E}(\mathbf{x})$ and $\mathbf{Cov}(\mathbf{x})$ in this setting?

Answer: $\mathbb{E}(x) = \mathbb{E}_x\left[\prod_{i=1}^{D} \mu_i^{x_i}(1-\mu_i)^{1-x_i}\right] = \cancel{\log\left[\sum \mu_i^{x_i}(1-\mu_i)^{1-x_i}\right]}$

$\boxed{\mathbb{E}[X] = \mu}$ $= \log\left(\mu_i^{x_i}\right) + \log\left((1-\mu_i)^{1-x_i}\right) = x_i \log(\mu_i) + (1-x_i)\log(1-\mu_i)$

$= E_x[x_i \log(\mu_i)] + E_x[1-x_i \log(1-\mu_i)] = \log(\mu_i) E_x[x_i]$

μ $(1-\mu)$

$COV(X) = \cancel{\text{Diag}}$ $\text{Diag}[\mu_i(1-\mu_i)] = I\,(\mu_i(1-\mu_i))$

3 pts.

b) Consider a Bernoulli mixture model for binary random vectors with K mixture components, i.e.

prior, weights

likelihood

$$p(\mathbf{x};\, (\boldsymbol{\mu}_k)_k, \mathbf{w}) = \sum_{k=1}^{K} w_k p(\mathbf{x};\, \boldsymbol{\mu}_k)$$

with weights $\mathbf{w} = (w_1, \ldots, w_K)^T$ and Bernoulli parameters $(\boldsymbol{\mu}_k)_k = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$.

(Side remark, which is not relevant to solve the question: The number of mixture components K cannot be larger than the dimension D of the random vector to ensure identifiability.)

What are natural constraints on $\mathbf{w}$ to obtain a proper probability distribution $p(\mathbf{x};\ (\boldsymbol{\mu}_k)_k, \mathbf{w})$?
Answer:

2 pts.

c) Show that

$$\mathbb{E}(\mathbf{x}) = \sum_k w_k \boldsymbol{\mu}_k \qquad \text{and}$$

$$(\mathbf{Cov}(\mathbf{x}))_{i,j} = \begin{cases} \sum_k w_k \mu_{ki} - \mathbb{E}(\mathbf{x}_i)^2 & \text{if } i = j \\ \sum_k w_k \mu_{ki} \mu_{kj} - \mathbb{E}(\mathbf{x}_i)\mathbb{E}(\mathbf{x}_j) & \text{if } i \neq j \end{cases}$$

(the elements of the covariance matrix)

where μ_{ki} is the i-th component of $\boldsymbol{\mu}_k$.

Hints:

1. You do not need to include the last terms, $-\mathbb{E}(\mathbf{x}_i)^2$ and $-\mathbb{E}(\mathbf{x}_i)\mathbb{E}(\mathbf{x}_j)$, in your derivation, since these follow from the relation

$$\mathbf{Cov}(\mathbf{x}) = \mathbb{E}(\mathbf{x}\mathbf{x}^T) - \mathbb{E}(\mathbf{x})\mathbb{E}(\mathbf{x})^T.$$

   Thus, you need to focus only on $\mathbb{E}(\mathbf{x}\mathbf{x}^T)$.
2. Consider $\mathbb{E}(\mathbf{x}_i\mathbf{x}_j)$ separately for the cases $i = j$ and $i \neq j$. Use the fact that $\mathbb{E}(\mathbf{x}_i\mathbf{x}_j) = P(\mathbf{x}_i = 1 \wedge \mathbf{x}_j = 1)$, since $\mathbf{x}$ is a binary vector.

What is the main (qualitative) difference between the covariance derived in a) and the one for the mixture model?

Answer:

5 pts.

d) In order to derive an EM-method to determine the mixture of Bernoulli paramaters, we introduce a latent binary vector $\mathbf{z} = (z_1, \dots, z_K)^T \in \{0,1\}^K$ linked with $\mathbf{x}$, such that $\sum_k z_k = 1$ and $z_k = 1$ if $\mathbf{x}$ is generated by the k-th mixture component. Hence, we have

$$p(\mathbf{x} \mid \mathbf{z}) = \prod_{k=1}^{K} p(\mathbf{x};\, \boldsymbol{\mu}_k)^{z_k} \qquad \text{and} \qquad p(\mathbf{z}) = \prod_{k=1}^{K} w_k^{z_k}.$$

Derive $p(\mathbf{x})$ by marginalizing over $\mathbf{z}$. Answer:

$$p(z|x) = \frac{P(x|z)^{z_k} \cdot P(z)}{p(x)} \qquad p(x) = \int p(x|z)^{z_k} \cdot p(z)$$

3 pts.

e) M-step: let $\mathbf{X} := (\mathbf{x}_n)_n \in \mathbb{R}^{D \times N}$ be a sequence of N i.i.d. samples drawn from the mixture of Bernoulli distribution. We have a corresponding matrix of the latent variables $\mathbf{Z} := (\mathbf{z}_n)_n \in \mathbb{R}^{K \times N}$.

Write down the joint log-likelihood function $\ln p(\mathbf{X}, \mathbf{Z};\, (\boldsymbol{\mu}_k)_k, \mathbf{w})$ and derive the M-step by maximizing the joint log-likelihood over the unknown parameters $(\boldsymbol{\mu}_k)_k$ and $\mathbf{w}$. Hint: don't forget to introduce a Lagrange multiplier for the constraint on $\mathbf{w}$.

Answer:

6 pts.

f) E-step: derive the expectation of the joint log-likelihood function with respect to $\mathbf{Z}$. Hints:

1. Derive $p(\mathbf{Z} \mid \mathbf{X})$ first via $p(\mathbf{z})$ and $p(\mathbf{x} \mid \dot{\mathbf{z}})$ given above.
2. Consider $\mathbb{E}(z_{nk}) = p(z_{nk} = 1)$, since $z_{nk} \in \{0, 1\}$.
3. Observe that z_{nk} appears only linearly in the joint log-likelihood function.

Answer:

$$p(z|x) = \frac{p(x|z) \cdot p(z)}{p(x)} = \frac{\prod_{k}^{K} p(x|\mu_k)^{z_k} \cdot \prod_{k}^{k} \omega_k^{z_k}}{p(x)}$$

$$\prod_{k}^{K} p(x|\mu_k)^{z_k}$$

log-likelihood fxn:

samples mixtures

$$\cancel{\log p(x|\mu)} \quad \log p(x, z|\mu_k) = \sum_{i}^{N} \sum_{j}^{K} z_{ij} \left(\log \omega_k \right) +$$

$$\sum_{k}^{D} (x_{nk} \ln \mu$$

6 pts.

This page has been intentionally left blank.

$S^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$

Supplementary Sheet

$E[S^2]$

Primal form.

$$L(w, w_0, \xi) = \frac{1}{2}w^2 + \frac{C}{2}\sum_i^N \xi_i^2 - \sum_i^N \alpha_i(y_i w_i x_i + w_0 - 1 + \xi_i) - \sum_i^N \lambda_i \xi_i = 0$$

$$\frac{\partial L}{\partial w} = w - \sum_i^N \alpha_i y_i x_i = 0 \qquad w = \sum_i^N \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial w_0} = \sum_i^N \alpha_i = 0$$

$$\frac{\partial L}{\partial \xi} = C\sum_i^N \xi_i - \sum_i^N \lambda_i - \sum_i^N \alpha_i = 0 \rightarrow C\sum_i^N \xi_i = \sum_i^N \lambda_i + \sum_i^N \alpha_i = 0 \Rightarrow \sum_i^N \xi_i = \frac{1}{C}\left(\sum_i^N \lambda_i + \sum_i^N \alpha_i\right)$$

~~Dual~~ Dual ~~Primal~~ form :

$$\max_{\alpha} \frac{1}{2}\left(\sum_i^N \alpha_i \cancel{y_i} x_i\right)\left(\sum_j^N \alpha_j y_j x_j\right) + \frac{C}{2}\sum_i^N \xi_i^2 - \sum_i^N \alpha_i \left(y_i \left(\sum_j^N \alpha_j y_j x_j\right) x_i - \sum_i^N w_0 \alpha_i + \sum_i^N \alpha_i\right.$$

$$- \sum_i^N \alpha_i \xi_i - \sum_i^N \lambda_i \xi_i = 0$$

$$\Rightarrow -\frac{1}{2}\sum_i\sum_j \alpha_i \alpha_j y_i y_j x_i x_j + \frac{C}{2}\sum_i^N \xi_i^2 - \sum_i^N w_0 \alpha_i + \sum_i^N \alpha_i - \sum_i^N \alpha_i \xi_i - \sum_i^N \lambda_i \xi_i = 0$$