PRINCE OF SONGKLA UNIVERSITY
FACULTY OF ENIGNEERING

Final Examination: Semester I — Academic Year: 2011

Date: August 2nd, 2011 — Time: 9:00-11:00 (2 hours)

Subject: 241-588 — Room: A305

SPECIAL TOPICS IN COMPUTER CONTROL ENGINEERING I
(INTRODUCTION TO MACHINE LEARNING)

Instructions:

1. Closed book
2. Write the answer in paper.
3. Calculate & computer notebook not allowed.

Name:_____

Student ID: _____

241-588 Introduction to Machine Learning

Semester 1, Year 2554 Mid-term Exam & Solution

By Anant Choksuriwong

Question	Topic	Max. score	Score
1	Short questions	20	
2	Bayes Optimal Classification	15	
3	Logistic Regression	18	
4	Regression	16	
5	SVM	16	
6	Boosting	15	
	Total	100	

1 Short Questions [20 pts]

Are the following statements True/False? Explain your reasoning in only 1 sentence.

1. Density estimation (using say, the kernel density estimator) can be used to perform classification.

2. The correspondence between logistic regression and Gaussian Naïve Bayes (with identity class covariances) means that there is a one-to-one correspondence between the parameters of the two classifiers.

3. The training error of 1-NN classifier is 0.

4. As the number of data points grows to infinity, the MAP estimate approaches the MLE estimate for all possible priors. In other words, given enough data, the choice of prior is irrelevant.

5. Cross validation can be used to select the number of iterations in boosting; this procedure may help reduce overfitting.

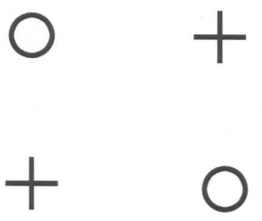
6. The kernel density estimator is equivalent to performing kernel regression with the value $Y_i = \frac{1}{n}$ at each point X_i in the original data set.

7. We learn a classifier f by boosting weak learners h. The functional form of f's decision boundary is the same as h's, but with different parameters. (e.g., if h was a linear classifier, then f is also a linear classifier).

8. The depth of a learned decision tree can be larger than the number of training examples used to create the tree.

For the following problems, circle the correct answers:

1. Consider the following data set:



Circle all of the classifiers that will achieve zero training error on this data set. (You may circle more than one.)

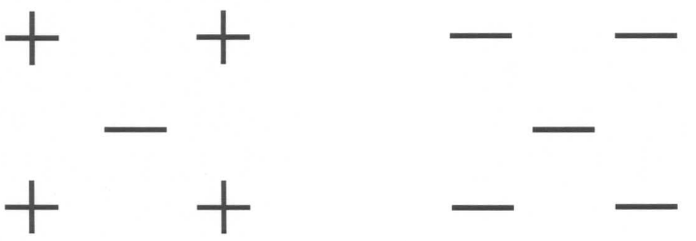
(a) Logistic regression

(b) SVM (quadratic kernel)

(c) Depth-2 ID3 decision trees

(d) 3-NN classifier

2. For the following dataset, circle the classifier which has larger Leave-One-Out Cross-validation error.



a) 1-NN

b) 3-NN

2 Bayes Optimal Classification [15 pts]

In classification, the loss function we usually want to minimize is the 0/1 loss:

$$\ell(f(x), y) = \mathbf{1}\{f(x) \neq y\}$$

where $f(x), y \in \{0, 1\}$ (i.e., binary classification). In this problem we will consider the effect of using an asymmetric loss function:

$$\ell_{\alpha,\beta}(f(x), y) = \alpha\mathbf{1}\{f(x) = 1, y = 0\} + \beta\mathbf{1}\{f(x) = 0, y = 1\}$$

Under this loss function, the two types of errors receive different weights, determined by $\alpha, \beta > 0$.

1. **[4 pts]** Determine the Bayes optimal classifier, i.e. the classifier that achieves minimum risk assuming $P(x, y)$ is known, for the loss $\ell_{\alpha,\beta}$ where $\alpha, \beta > 0$.

2. **[3 pts]** Suppose that the class $y = 0$ is extremely uncommon (i.e., $P(y = 0)$ is small). This means that the classifier $f(x) = 1$ for all x will have good risk. We may try to put the two classes on even footing by considering the risk:

   $$R = P(f(x) = 1|y = 0) + P(f(x) = 0|y = 1)$$

   Show how this risk is equivalent to choosing a certain α, β and minimizing the risk where the loss function is $\ell_{\alpha,\beta}$.

3. **[4 pts]** Consider the following classification problem. I first choose the label $Y \sim \text{Bernoulli}(\frac{1}{2})$, which is 1 with probability $\frac{1}{2}$. If $Y = 1$, then $X \sim \text{Bernoulli}(p)$; otherwise, $X \sim \text{Bernoulli}(q)$. Assume that $p > q$. What is the Bayes optimal classifier, and what is its risk?

4. **[4 pts]** Now consider the regular 0/1 loss ℓ, and assume that $P(y = 0) = P(y = 1) = 1/2$. Also, assume that the class-conditional densities are Gaussian with mean μ_0 and co-variance Σ_0 under class 0, and mean μ_1 and co-variance Σ_1 under class 1. Further, assume that $\mu_0 = \mu_1$.

   For the following case, draw contours of the level sets of the class conditional densities and label them with $p(x|y = 0)$ and $p(x|y = 1)$. Also, draw the decision boundaries obtained using the Bayes optimal classifier in each case and indicate the regions where the classifier will predict class 0 and where it will predict class 1.

$$\Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$

3 Logistic Regression [18 pts]

We consider here a discriminative approach for solving the classification problem illustrated in Figure 1.

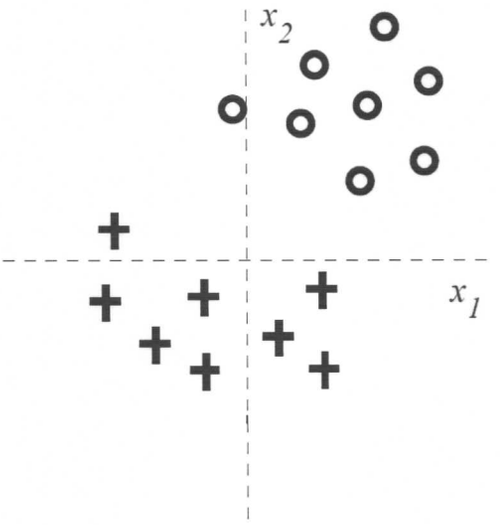


Figure 1: The 2-dimensional labeled training set, where '+' corresponds to class y=1 and 'O' corresponds to class $y = 0$.

1. We attempt to solve the binary classification task depicted in Figure 1 with the simple linear logistic regression model

$$P(y = 1|\vec{x}, \vec{w}) = g(w_0 + w_1x_1 + w_2x_2) = \frac{1}{1 + exp(-w_0 - w_1x_1 - w_2x_2)}.$$

Notice that the training data can be separated with *zero* training error with a linear separator.

Consider training *regularized* linear logistic regression models where we try to maximize

$$\sum_{i=1}^{n} \log\left(P(y_i|x_i, w_0, w_1, w_2)\right) - Cw_j^2$$

for very large C. The regularization penalties used in penalized conditional log-likelihood estimation are $-Cw_j^2$, where $j = \{0, 1, 2\}$. In other words, only one of the parameters is regularized in each case. Given the training data in Figure 1, how does the training error change with regularization of each parameter w_j? State whether the training error increases or stays the same (zero) for each w_j for very large C. Provide a brief justification for each of your answers.

(a) By regularizing w_2 **[2 pts]**

(b) By regularizing w_1 **[2 pts]**

(c) By regularizing w_0 **[2 pts]**

2. If we change the form of regularization to L1-norm (absolute value) and regularize w_1 and w_2 only (but not w_0), we get the following penalized log-likelihood

$$\sum_{i=1}^{n} \log P(y_i|x_i, w_0, w_1, w_2) - C(|w_1| + |w_2|).$$

Consider again the problem in Figure 1 and the same linear logistic regression model $P(y = 1|\vec{x}, \vec{w}) = g(w_0 + w_1x_1 + w_2x_2)$.

(a) **[3 pts]** As we increase the regularization parameter C which of the following scenarios do you expect to observe? (Choose only one) Briefly explain your choice:

( ) First w_1 will become 0, then w_2.

( ) First w_2 will become 0, then w_1.

( ) w_1 and w_2 will become zero simultaneously.

( ) None of the weights will become exactly zero, only smaller as C increases.

(b) **[3 pts]** For very large C, with the same L1-norm regularization for w_1 and w_2 as above, which value(s) do you expect w_0 to take? Explain briefly. (Note that the number of points from each class is the same.) (You can give a range of values for w_0 if you deem necessary).

(c) **[3 pts]** Assume that we obtain more data points from the '+' class that corresponds to y=1 so that the class labels become unbalanced. Again for very large C, with the same L1-norm regularization for w_1 and w_2 as above, which value(s) do you expect w_0 to take? Explain briefly. (You can give a range of values for w_0 if you deem necessary).

4 Kernel regression [16 pts]

Now lets consider the non-parametric kernel regression setting. In this problem, you will investigate univariate locally linear regression where the estimator is of the form:

$$\widehat{f}(x) = \beta_1 + \beta_2 x$$

and the solution for parameter vector $\beta = [\beta_1 \ \beta_2]$ is obtained by minimizing the weighted least square error:

$$J(\beta_1, \beta_2) = \sum_{i=1}^{n} W_i(x)(Y_i - \beta_1 - \beta_2 X_i)^2 \quad \text{where} \quad W_i(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)},$$

where K is a kernel with bandwidth h. Observe that the weighted least squares error can be expressed in matrix form as

$$J(\beta_1, \beta_2) = (Y - A\beta)^T W (Y - A\beta),$$

where Y is a vector of n labels in the training example, W is a $n \times n$ diagonal matrix with weight of each training example on the diagonal, and

$$A = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \cdots & \\ 1 & X_n \end{bmatrix}$$

1. [**4 pts**] Derive an expression in matrix form for the solution vector $\hat{\beta}$ that minimizes the weighted least square.

2. [**3 pts**] When is the above solution unique?

3. [**3 pts**] If the solution is not unique, one approach is to optimize the objective function J using gradient descent. Write the update equation for gradient descent in this case. Note: Your answer must be expressed in terms of the matrices defined above.

4. [**3 pts**] Can you identify the signal plus noise model under which maximizing the likelihood (MLE) corresponds to the weighted least squares formulation mentioned above?

5. [**3 pts**] Why is the above setting non-parametric? Mention one advantage and one disadvantage of nonparametric techniques over parametric techniques.

5 SVM [16 pts]

5.1 L2 SVM

Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ be a set of l training pairs of feature vectors and labels. We consider binary classification, and assume $y_i \in \{-1, +1\}\ \forall i$. The following is the primal formulation of L2 SVM, a variant of the standard SVM obtained by squaring the hinge loss:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\mathbf{w}^\top\mathbf{w} + \frac{C}{2}\sum_{i=1}^{l}\xi_i^2 \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i \in \{1, \ldots, l\}, \\ & \xi_i \geq 0, \quad i \in \{1, \ldots, l\}. \end{aligned}$$

1. **[4 pts]** Show that removing the last set of constraints $\{\xi_i \geq 0\ \forall i\}$ does not change the optimal solution to the primal problem.

2. **[3 pts]** After removing the last set of constraints, we get a simpler problem:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\mathbf{w}^\top\mathbf{w} + \frac{C}{2}\sum_{i=1}^{l}\xi_i^2 \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i \in \{1, \ldots, l\}. \end{aligned} \tag{1}$$

   Give the Lagrangian of (1).

3. **[6 pts]** Derive the dual of (1). How is it different from the dual of the standard SVM with the hinge loss?

5.2 Leave-one-out Error and Support Vectors

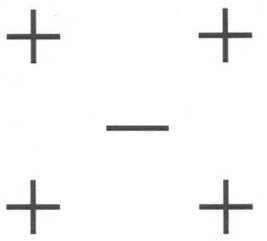
[3 pts] Consider the standard two-class SVM with the hinge loss. Argue that under a given value of C,

$$\text{LOO error} \leq \frac{\#\text{SVs}}{l},$$

where l is the size of the training data and #SVs is the number of support vectors obtained by training SVM on the entire set of training data.

6 Boosting [15 pts]

1. Consider training a boosting classifier using decision stumps on the following data set:



(a) **[3 pts]** Which examples will have their weights increased at the end of the first iteration? Circle them.

(b) **[3 pts]** How many iterations will it take to achieve zero training error? Explain.

(c) **[3 pts]** Can you add one more example to the training set so that boosting will achieve zero training error in two steps? If not, explain why.

2. **[2 pts]** Why do we want to use "weak" learners when boosting?

3. **[4 pts]** Suppose AdaBoost is run on m training examples, and suppose on each round that the weighted training error ϵ_t of the t^{th} weak hypothesis is at most $1/2 - \gamma$, for some number $\gamma > 0$. After how many iterations, T, will the combined hypothesis H be consistent with the m training examples, i.e., achieves zero training error? Your answer should only be expressed in terms of m and γ. (Hint: What is the training error when 1 example is misclassified?)