Practice Midterm Exam

EECS 545: Machine Learning
Winter, 2016

Name:

UM uniqname:

- **Closed book. Three sheets of paper of notes are allowed. No computers, cell phones or calculators.**
  - Showing your work makes partial credit possible.
    If you write nothing at all, it's hard to justify any score but zero.
  - Feel free to use the backs of the sheets for scratch paper.
  - Write clearly. If we can't read your writing, it will be marked wrong.
- This course operates under the rules of the College of Engineering Honor Code. Your signature endorses the pledge below. **After** you finish your exam, please sign below:
  I have neither given nor received aid on this examination, nor have I concealed any violations of the Honor Code.

Problem 1 (True/False). Are the following statements true or false? (No need for explanations unless you feel the question is ambiguous and want to justify your answer).

1. The error on the training set is a better estimate of the generalization error than the error on the test set.

2. The perceptron algorithm finds the maximum margin classifier if the data is linearly separable.

3. Bayesian reasoning is popular since it avoids the need to explicitly specify a prior distribution.

4. Assume we have trained a model for linear disciminant analysis, and we obtained parameters Σ, the covariance matrix, and μ_1, μ_2, the class means. We learned in class that the decision boundary between classes $c = 0$ and $c = 1$, i.e. the set $\{\mathbf{x} : P(y = c|\mathbf{x}, \Sigma, \mu_1, \mu_2) = 0.5\}$, is linear in the input space. But it is not linear at thresholds other than 0.5; for example, the set $\{\mathbf{x} : P(y = c|\mathbf{x}, \Sigma, \mu_1, \mu_2) = 0.9\}$ is not an affine subspace.

5. Locally-weighted linear regression can produce nonlinear fits to the data.

6. The specification of a probabilistic discriminative model can often be interpreted as a method for creating new, "fake" data.

7. Gaussian Discriminant Analysis as an approach to classification cannot be **applied** if the true class-conditional density for each class is *not* Gaussian.

8. Linear Regression can only be applied when the target values are binary or discrete.

9. The soft-margin SVM tends to have larger margin when the parameter C increases.

10. To solve

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \quad s.t. \quad \begin{array}{l} \sum_{i=1}^n x_i = 1 \\ \mathbf{x}_i \geq 0, \ \forall i \end{array}$$

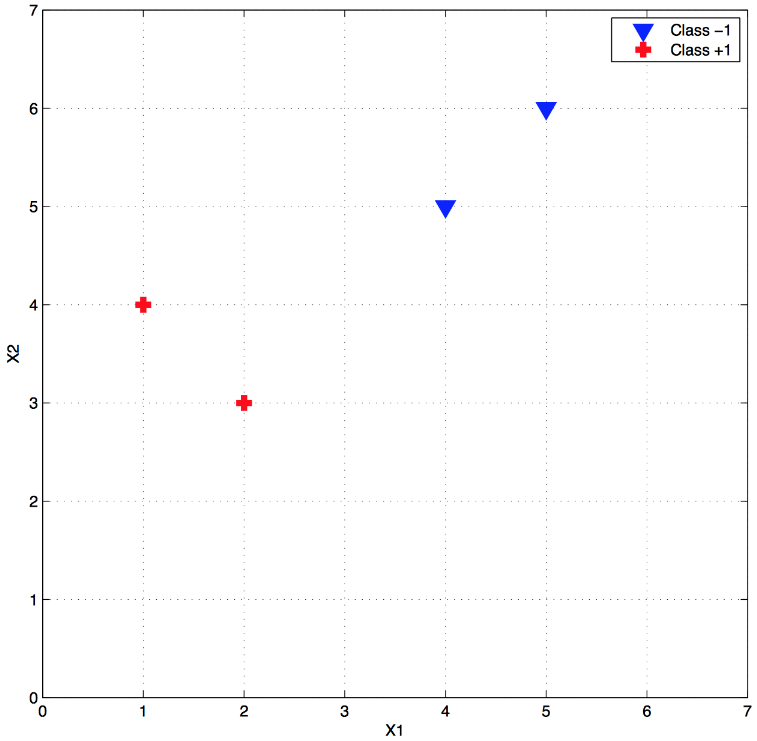
The Lagrangian would be $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \nu) = f(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{x} + \nu(\mathbf{1}^T\mathbf{x} - 1)$ where $\boldsymbol{\lambda} \in \mathbb{R}^n$, $\lambda_i \geq 0$, $\nu \in \mathbb{R}$, and $\mathbf{1}$ is a vector of length n of all 1s.

Problem 2 (Probability). For data D and hypothesis H, say whether or not the following equations must always be true.

a $\sum_h P(H = h|D = d) = 1$

b $\sum_h P(D = d|H = h) = 1$

c $\sum_h P(D = d|H = h)P(H = h) = 1$



Problem 3 (SVM). 1. In class we learnt that SVM can be used to classify linearly inseparable data by transforming it to a higher dimensional space with a kernel $K(x, z) = \phi(x)^T \phi(z)$, where $\phi(x)$ is a feature mapping. Let K_1 and K_2 be $R^n \times R^n$ kernels, K_3 be a $R^d \times R^d$ kernel and $c \in R^+$ be a positive constant. $\phi_1 : R^n \to R^d$, $\phi_2 : R^n \to R^d$, and $\phi_3 : R^d \to R^d$ are feature mappings of K_1, K_2 and K_3 respectively. Explain how to use ϕ_1 and ϕ_2 to obtain the following kernels.

a $K(x, z) = cK_1(x, z)$

b $K(x, z) = K_1(x, z)K_2(x, z)$

2. Support vector machines learn a decision boundary leading to the largest margin from both classes. You are training SVM (with no slack) on a tiny dataset with 4 points shown in Figure above. This dataset consists of two examples with class label -1 (denoted with plus), and two examples with class label $+1$ (denoted with triangles).

   (a) Find the weight vector w and bias b. What is the equation corresponding to the decision boundary?

(b) Circle the support vectors and draw the decision boundary in Figure above.

Problem 4 (Coin Flips and Pseudocounts). Suppose we flip a (not necessarily fair) coin N times and wish to estimate its bias θ after observing X heads. We endow θ with a Beta prior. Mathematically, our model is

$$\begin{aligned} \theta &\sim \mathrm{Beta}(a, b) \\ X &\sim \mathrm{Binomial}(N, \theta) \end{aligned}$$

Part A. Derive the maximum likelihood estimate $\hat{\theta}_{ML}$ of the coin's bias? Show your work.

Part B. Write down the corresponding MAP estimate $\hat{\theta}_{MAP}$. No need to show your work.

Problem 5 (Irrelevant Features with Naive Bayes). In this exercise, we consider words that are *nondiscriminative* for document classification (such as 'the', 'and', etc.) and analyze their impact on the decision made by Naive Bayes in several settings.

Let $x_{dw} = 1$ if word w occurs in document d and $x_{dw} = 0$ otherwise. Let the vocabulary size be W, and let θ_{cw} be the estimated probability $P(x_{dw} = 1|c)$ that word w occurs in documents of class c. Recall that the joint likelihood for Naive Bayes is

$$P(\mathbf{x}_d, c|\theta) = P(\mathbf{x}_d|c, \theta) = P(c) \prod_{w=1}^{W} P(x_{dw}|\theta_{cw})$$

where $P(c)$ specifies the class priors, and $\mathbf{x}_d = (\mathbf{x}_{d1}, \ldots, \mathbf{x}_{dW})$ is a document.

Part A. Here, we show that Naive Bayes is a linear classifier. Define the new parameter vector

$$\beta_c = \left(\log \frac{\theta_{c1}}{1 - \theta_{c1}}, \cdots, \log \frac{\theta_{cW}}{1 - \theta_{cW}}, \sum_{w=1}^{W} \log(1 - \theta_{cw}) \right)^T$$

and let $\phi(\mathbf{x}_d) = (x_{d1}, \ldots, x_{dW}, 1)^T$. Show that $\log P(\mathbf{x}_d|c, \theta) = \phi(\mathbf{x}_d)^T \beta_c$.

Part B. Suppose there are only two possible document classes c_A and c_B, and assume a uniform class prior $\pi_A = \pi_B = 0.5$. and find an expression for the log posterior odds ratio R, shown below, in terms of the features $\phi(\mathbf{x}_d)$ and the parameters β_1 and β_2.

$$R = \log \frac{P(c_A|\mathbf{x}_d)}{P(c_B|\mathbf{x}_d)}$$

Part C. Intuitively, words that occur in both classes are not very *discriminative*, and therefore should not affect our beliefs about the class label. State the conditions under which the presence or absence of a particular word w in a test document will have no effect on the class posterior (such a word will effectively be ignored by the classifier).

<u>Part D.</u> Consider a set of documents $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with labels $\mathcal{Y} = \{y_1, \ldots, y_n\}$. Suppose a particular word w always occurs in every document, regardless of class. Let there be N_A and N_B documents in classes A and B respectively, where $N_A \neq N_B$ (class imbalance). If we estimate the parameters θ_{cw} with the posterior mean under a uniform $\text{Beta}(1, 1)$ prior after observing data $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, will word w be ignored by our classifier?

Problem 6 (Convexity). Let $J(\boldsymbol{\theta})$ be a twice-differentiable function such that

$$\nabla^2 J(\boldsymbol{\theta}) \preceq B$$

i.e., $B - \nabla^2 J(\boldsymbol{\theta})$ is positive semi-definite for some fixed positive definite matrix B (independent of $\boldsymbol{\theta}$). Show that given a fixed value $\boldsymbol{\theta}^{(t)}$, the function

$$J_t(\boldsymbol{\theta}) = J(\boldsymbol{\theta}^{(t)}) + \nabla J(\boldsymbol{\theta}^{(t)})^T(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})^T B(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})$$

is a majorizing function of $J(\boldsymbol{\theta})$; i.e., for all $\boldsymbol{\theta}$, $J_t(\boldsymbol{\theta}) \geq J(\boldsymbol{\theta})$, and $J_t(\boldsymbol{\theta}^{(t)}) = J(\boldsymbol{\theta}^{(t)})$.

Hint: A twice continuously differentiable function f admits the quadratic expansion

$$f(\mathbf{x}) = f(\mathbf{y}) + \langle \triangledown f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{1}{2}\langle \mathbf{x} - \mathbf{y}, \triangledown^2 f(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))(\mathbf{x} - \mathbf{y}) \rangle$$

for some $t \in (0, 1)$.

Problem 7 (Logistic Regression). Assume we have a training dataset that is linearly separable. Assume we train a logistic regression on this dataset with fixed parameters (we use the standard sigmoid function). Our logistic regression function predicts a probability for each new example, but assume we convert this to a classifier by thresholding the probability at $p \geq 0.5$ and $p < 0.5$. Question: if we measured this error on the training set, is it guaranteed that this error is zero?

Either prove that it does have zero training error or propose a dataset where the logistic regression returns a classifier which has non-zero training error.