Practice Exam
EECS 545: Machine Learning
Fall, 2017

Name:

UM uniqname:

- **Closed book. Two 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. Bayesian reasoning is popular since it avoids the need to explicitly specify a prior distribution.

3. Assume we have trained a model for linear discriminant 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 $\{\mathbf{x} : P(y = c|\mathbf{x}, \Sigma, \mu_0, \mu_1) = 0.5\}$, 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.

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

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

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

7. The soft-margin SVM tends to have larger margin when the parameter C increases.
8. (1 pt) (True/False) The optimization problem for hard-margin SVM always has at least one feasible solution for any training dataset.

9. (1 pt) (True/False) In the least squares regression problem $\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^{\mathbf{2}}$, there may be more than one $\mathbf{w}$ that minimizes this objective. As a result, there may be more than one correct prediction $\widehat{\mathbf{w}}^T\mathbf{x}$.

10. (1 pt) (True/False) The dual norm of a norm $\|.\|$ is denoted $\|.\|_*$ and is defined as

$$\|\mathbf{x}\|_* = \max_{\mathbf{z}:\|\mathbf{z}\|\leq 1} \mathbf{x}^T\mathbf{z}.$$

In the optimization problem, let the norm be the l^p norm, i.e. for $p \geq 1$, $\|\mathbf{x}\| = \|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$, therefore the Lagrangian is $L(\mathbf{z}, \lambda) = -\mathbf{x}^T\mathbf{z} + \lambda\left(\sum_{i=1}^{n} |z_i|^p - 1\right)$.

11. The principal eigenvector of PCA, i.e.,

$$\arg \max_{u_1:\ u_1^T u_1 = 1} \sum_{i=1}^{n} (u_1^T (x_i - \bar{x}))^2$$

is always unique.

12. We need labels to apply k-means clustering.

Problem 2 (Kernels and 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, and k_3 be $R^n \times R^n$ kernels and $c_1, c_2 \in R^+$ be positive constants. $\phi_1 : R^n \to R^d$, $\phi_2 : R^n \to R^d$, and $\phi_3 : R^n \to R^d$ 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.

   1 $k(x, z) = c_1 k_1(x, z)$

   b $k(x, z) = c_1 k_1(x, z) + c_2 k_2(x, z)$

2. Consider a generic soft-margin SVM optimization problem:

$$
\begin{aligned}
\min_{w,b,s_1,\ldots,s_n} \quad & \frac{1}{2}\|w\|_2^2 + C\frac{1}{n}\sum_{i=1}^{n} s_i \\
\text{subject to} \quad & y_i(x_i^T w + b) \geq 1 - s_i \quad \text{for } i = 1, \ldots, n \\
& s_i \geq 0 \quad \text{for } i = 1, \ldots, n \,,
\end{aligned}
$$

   Suppose that we add the constraints $s_i = s_j \;\; \forall ij$, in the optimization problem to make every slack variable equal to each other. Transform the constrained problem into an unconstrained optimization problem by eliminating $s_1, \ldots, s_n$.

Problem 3 (One-Class Support Vector Machine). This problem will explore an SVM-like algorithm called the one-class SVM. Consider a classification problem where there are two classes, but we only have training data from one of the classes. Let $\mathbf{x_1}, \ldots, \mathbf{x_n}$ denote the training data from this class. The goal is to design a good classier even though we have no data from the other class. This problem is often referred to as one-class classification, anomaly detection, or novelty detection (the unobserved class is viewed as an anomaly or novelty).
Let $L(t) = \max\{0, 1-t\}$ be the hinge loss. Consider the optimization problem

$$\min_{\mathbf{w}} \frac{\lambda}{2} \|\mathbf{w}\|^{\mathbf{2}} + \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{L}(\mathbf{w}^{\mathbf{T}} \mathbf{x_i}), \tag{1}$$

where $\lambda > 0$ is fixed. The solution $\mathbf{w}$ defines an anomaly detector, called the *one-class support vector machine* (OC-SVM), by the function

$$f(\mathbf{x}) = \mathbf{sign}\{\mathbf{w}^{\mathbf{T}}\mathbf{x} - \mathbf{1}\},$$

where a prediction of $+1$ corresponds to the observed class, and -1 to the unobserved class. At first glance, it may not be clear why this is a good approach to one-class classification. Below, when we kernelize the algorithm, the utility of this classier will be more apparent.

a. (5 points) Rewrite the above optimization problem as a quadratic program in the variables $\mathbf{w}$ and $\zeta_1, \ldots, \zeta_n$, where ζ_i are slack variables.

b. (5 points) Derive the dual optimization problem to the quadratic program from part a. You do not need to explain how to solve the dual.

c. (5points) Explain how to kernelize the OC-SVM. In the case of the Gaussian kernel, provide an intuitive interpretation of classier.

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

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

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.

<u>Part A.</u> 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).

Part D. 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

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

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.