Practice Midterm Exam

EECS 545-002: Machine Learning

Fall, 2017

Most of these problems are former exam problems, and all of them should help you study for the exam. If you want to simulate an exam, try to work problems in 3 hours. Keep in mind that you will be allowed two cheat sheets (front and back, handwritten, regular printer paper) for the exam. Also remember that there are topics covered in class that are not represented below, although I tried to select problems that provide reasonable coverage of all concepts.

1 Short Answers: General ML Topics (14 points)

For the "True/False" questions, please circle the correct answer.

a. (1 pt) (True/False) If the training data are separable, it is always better to use a hard-margin SVM since it assumes linear separability of the training data.

b. (1 pt) (True/False) In terms of the optimization problem each learning algorithm solves, the loss function is the main difference between between L2-regularized logistic regression and Support Vector Machine.

c. (1 pt) (True/False) Locally-weighted linear regression can produce nonlinear fits to the data.

d. (1 pt) (True/False) The core assumption of Naive Bayes classifiers is that all observed variables (features) are statistically independent.

e. (1 pt) (True/False) In PCA, let $\theta_1, \ldots, \theta_n \in \mathcal{R}^k$ be the reduced-dimensionality representation of training data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{R}^d$. Then the principle components are uncorrelated i.e. that the sample covariance matrix of $\theta_1, \ldots, \theta_n$ is diagonal.

f. (1 pt) (True/False) In gradient descent algorithm, considering large step size implies fast convergence and small residual error.

g. (1 pt) (True/False) Let $\mathcal{H} = \{\mathbf{x} | x^{(1)} - 2x^{(2)} + 3x^{(3)} - 4 = 0\}$. The distance from $\mathcal{H}$ from $[1 \;\; 1 \;\; 1]^T$ is $1/\sqrt{14}$.

h. (1 pt) (True/False) In kernel PCA, we solve the eigenvalue problem $n\lambda\alpha = K\alpha$, where K is gram matrix $K = [k_{ij}]_{i,j=1}^n$ and α is the corresponding eigenvector of K.

i. (1 pt) (True/False) The optimization problem for hard-margin SVM always has at least one feasible solution for any training dataset.

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

k. (1 pt) (True/False) In the optimal soft-margin SVM, the support vectors are only points on the margin and within the margin.

l. (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 $L(\mathbf{z}, \lambda) = -\mathbf{x}^T\mathbf{z} + \lambda\left(\sum_{i=1}^n |z_i|^p - 1\right)$.

m. (2 pts) Suppose we are trying to predict how users rate movies. Each movie x is represented by a feature vector $\phi(x) \in \mathcal{R}^d$ with coordinates that pertain to properties of the movie. The ratings are values y. Given n already rated movies, we can use Ridge regression to predict ratings. The training criterion is

$$J(\theta) = \frac{1}{2}\|\theta\| + \frac{1}{n}\sum_{i=1}^n \left(y_i - \theta^T\phi(x_i)\right)^2/2. \tag{1}$$

where $\theta \in \mathcal{R}^d$. If $\hat{\theta}$ is the optimal setting of the parameters with respect to the above criterion, which of the following conditions must be true.

( ) $\lambda\hat{\theta} - \frac{1}{n}\sum_{i=1}^n \left(y_i - \theta^T\phi(x_i)\right) = 0$.

( ) $J(\hat{\theta}) \leq J(\theta)$ for all $\theta \in \mathcal{R}^d$ (for fixed λ).

( ) If we increase λ, the resulting $\hat{\theta}$ will decrease.

2 Logistic Regression with Label Noise (12 points)

When applying logistic regression, we ordinarily observe $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i \in \mathcal{R}^d$ and $y_i \in \{0, 1\}$. Supposed that instead of y_i however, we observe a possibly noise label $\tilde{y}_i$ In particular, suppose that if $y_i = 0$ then $\tilde{y}_i = 1$ with probability p_0, independent $\mathbf{x}_i$. Similarly, if $y_i = 1$, then $\tilde{y}_i = 0$ with probability p_1, independent of $\mathbf{x}_i$. Thus the training data $(\mathbf{x}_1, \tilde{y}_1), \ldots, (\mathbf{x}_n, \tilde{y}_n)$. In this problem you will develop the logistic regression classifier (considering p_0 and p_1) even though the true labels are not observed.

Let $\mathbf{w}$ and b denote the unknown parameters defining the linear classier. Let θ be a parameter vector including all of the unknowns, namely p_0, p_1, $\mathbf{w}$ and b. For convenience let's also denote $\alpha = [\mathbf{w^T b}]^\mathbf{T}$. Finally, let $\mathbf{X}$, Y, $\tilde{Y}$ denote the random variables associated with x_i, y_i, $\tilde{y}_i$.

a. (3 points) Recall $\eta(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x})$, and define $\tilde{\eta}(\mathbf{x}) = P(\tilde{Y} = 1|\mathbf{X} = \mathbf{x})$. Express $\tilde{\eta}(\mathbf{x})$ in terms of $\eta(\mathbf{x})$.

b. (2 points) Assume a logistic regression model for Y conditioned on $\mathbf{X}$. Use the result of part a. to state a corresponding model for $\tilde{Y}$ given $\mathbf{X}$.

c. (3 points) Write down the log-likelihood of θ given the observed training data.

d. (4 points) A natural state variable for this problem is simply the true label y_i. Using this state variable, the *complete data likelihood* is

$$\begin{aligned} &\prod_{i=1}^{n} P\{Y_i = y_i, \tilde{Y}_i = \tilde{y}_i | \mathbf{X_i} = \mathbf{x_i}, \theta\} \\ &= \prod_{i=1}^{n} P\{Y_i = y_i | \tilde{Y}_i = \tilde{y}_i, \mathbf{X}_i = \mathbf{x}_i, \theta\} \; P\{\tilde{Y}_i = \tilde{y}_i | \mathbf{X_i} = \mathbf{x_i}, \theta\} \end{aligned} \tag{2}$$

viewed as a function of θ. Express the complete-data log-likelihood as an explicit function of θ and the complete data.

3 Irrelevant Features with Naive Bayes (13 points)

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

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

where $\pi(c)$ specifies the class priors, and $x_d = (x_{d1}, \ldots, x_{dW})^T$ is a document. Define the new parameter vector

$$\beta_c = \left(\log \frac{\theta_{c1}}{1 - \theta_{c1}}, \ldots, \log \frac{\theta_{cW}}{1 - \theta_{cW}}, \sum_{w=1}^{W} \log(1 - \theta_{cw}) \right)^T, \tag{3}$$

and let $\phi(x_d) = (x_{d1}, \ldots, x_{dW}, 1)^T$.

a. (5 Points) Show that $P(x_d|c,\theta) = \phi(x_d)^T.\beta_c$.

b. (3 points) Is Naive Bayes a linear classier? (why).

c. (5 points) Suppose there are only two possible document classes c_A and c_B, and assume a uniform class prior $\pi_A = \pi_B = 0.5$. Find an expression for the log posterior odds ratio R, shown below, in terms of the features $\phi(x_d)$ and the parameters β_1, β_2.

$$R = \log \frac{P(c_A|x_d)}{P(c_B|x_d)}. \tag{4}$$

4 Fisher's Linear Discriminant (10 points)

Consider binary classification where the class labels are -1 and $+1$. Let the training data be $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, where $\mathbf{x}_i \in \mathcal{R}^d$. Let $\bar{\mathbf{x}}$ be the overall sample mean of the training patterns. Let μ_- and μ_+ denote the sample means of the two classes, where $\mu_-, \mu_+ \in \mathcal{R}^d$. Define the *between class scatter matrix*

$$S_b = (\mu_- - \mu_+)(\mu_- - \mu_+)^T,$$

and the *within class scatter matrix*

$$S_w = \sum_{k\in\{+,-\}} \sum_{i:y_i=k} (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T$$

Fisher's Linear Discriminant (FLD) seeks a linear classier with normal vector $\mathbf{w}$ obtained by solving

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}.$$

The offset b of the linear classier is determined by some other criterion after $\mathbf{w}$ has been determined- FLD has nothing to say about b.

a. (5 points) Determine the optimal $\mathbf{w}$ according to the FLD criterion.

b. (5 points) Justify FLD by interpreting the numerator and denominator of the objective function. Be as quantitative as possible. Hint: Think of FLD as a supervised version of PCA, but with just one principal component.

5 One-Class Support Vector Machine (15 points)

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

$$\min_{\mathbf{w}} \frac{\lambda}{2}\|\mathbf{w}\|^{\mathbf{2}} + \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{L}(\mathbf{w^T x_i}), \tag{5}$$

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

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

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.

6 k-medoids (6 points)

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be data points. The k-medoids algorithm is a clustering algorithm similar to k-means. Let $d(\mathbf{x}, \mathbf{x}')$ denote the distance between $\mathbf{x}$ and $\mathbf{x}'$, and let $\mathbf{m}_1, \ldots, \mathbf{m}_k$ denote cluster centroids. k-means iterates

- $C(i) = \underset{l=1,\ldots,k}{\arg\min}\ d(\mathbf{x}_i, \mathbf{m}_l)$
- $\mathbf{m}_l$ sample mean of data points $\mathbf{x}_i$ such that $C(i) = l$

In k-medoids, the second step is replaced by

$$\mathbf{m}_l = \underset{\mathbf{m} \in \{\mathbf{x}_i | C(i) = l\}}{\arg\min} \sum_{j: C(j) = l} d(\mathbf{x}_i, \mathbf{m}).$$

Note that in k-medoids, the cluster center must be a member of the cluster.

a. (4 points) Argue that k-medoids is more robust to outliers than k-means.

b. (2 points) Other than robustness, what is a significant advantage of k-medoids over k-means?