CS 229, Autumn 2010
Practice Midterm

Notes:

1. The midterm will have about 5-6 long questions, and about 8-10 short questions. Space will be provided on the actual midterm for you to write your answers.
2. The midterm is meant to be educational, and as such some questions could be quite challenging. Use your time wisely to answer as much as you can!
3. For additional practice, please see CS 229 extra problem sets available at

   http://see.stanford.edu/see/materials/aimlcs229/assignments.aspx

1. [**13 points**] **Generalized Linear Models**

   Recall that generalized linear models assume that the response variable y (conditioned on x) is distributed according to a member of the exponential family:

   $$p(y;\eta) = b(y)\exp(\eta T(y) - a(\eta)),$$

   where $\eta = \theta^T x$. For this problem, we will assume $\eta \in \mathbb{R}$.

   (a) [**10 points**] Given a training set $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$, the loglikelihood is given by

   $$\ell(\theta) = \sum_{i=1}^{m} \log p(y^{(i)} \mid x^{(i)}; \theta).$$

   Give a set of conditions on $b(y)$, $T(y)$, and $a(\eta)$ which ensure that the loglikelihood is a concave function of θ (and thus has a unique maximum). Your conditions must be reasonable, and should be as weak as possible. (E.g., the answer "any $b(y)$, $T(y)$, and $a(\eta)$ so that $\ell(\theta)$ is concave" is not reasonable. Similarly, overly narrow conditions, including ones that apply only to specific GLMs, are also not reasonable.)

   (b) [**3 points**] When the response variable is distributed according to a Normal distribution (with unit variance), we have $b(y) = \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}}$, $T(y) = y$, and $a(\eta) = \frac{\eta^2}{2}$. Verify that the condition(s) you gave in part (a) hold for this setting.

2. [**15 points**] **Bayesian linear regression**

   Consider Bayesian linear regression using a Gaussian prior on the parameters $\theta \in \mathbb{R}^{n+1}$. Thus, in our prior, $\theta \sim \mathcal{N}(\vec{0}, \tau^2 I_{n+1})$, where $\tau^2 \in \mathbb{R}$, and I_{n+1} is the $n+1$-by-$n+1$ identity matrix. Also let the conditional distribution of $y^{(i)}$ given $x^{(i)}$ and θ be $\mathcal{N}(\theta^T x^{(i)}, \sigma^2)$, as in our usual linear least-squares model.[1] Let a set of m IID training examples be given (with $x^{(i)} \in \mathbb{R}^{n+1}$). Recall that the MAP estimate of the parameters θ is given by:

$$\theta_{MAP} = \arg\max_{\theta} \left(\prod_{i=1}^{m} p(y^{(i)}|x^{(i)}, \theta) \right) p(\theta)$$

Find, in closed form, the MAP estimate of the parameters θ. For this problem, you should treat τ^2 and σ^2 as fixed, known, constants. [Hint: Your solution should involve deriving something that looks a bit like the Normal equations.]

3. **[18 points] Kernels**

   In this problem, you will prove that certain functions K give valid kernels. Be careful to justify every step in your proofs. Specifically, if you use a result proved either in the lecture notes or homeworks, be careful to state exactly which result you're using.

   (a) **[8 points]** Let $K(x,z)$ be a valid (Mercer) kernel over $\mathbb{R}^n \times \mathbb{R}^n$. Consider the function given by

   $$K_e(x,z) = \exp(K(x,z)).$$

   Show that K_e is a valid kernel. [Hint: There are many ways of proving this result, but you might find the following two facts useful: (i) The Taylor expansion of e^x is given by $e^x = \sum_{j=0}^{\infty} \frac{1}{j!} x^j$ (ii) If a sequence of non-negative numbers $a_i \geq 0$ has a limit $a = \lim_{i \to \infty} a_i$, then $a \geq 0$.]

   (b) **[8 points]** The Gaussian kernel is given by the function

   $$K(x,z) = e^{-\frac{||x-z||^2}{\sigma^2}},$$

   where $\sigma^2 > 0$ is some fixed, positive constant. We said in class that this is a valid kernel, but did not prove it. Prove that the Gaussian kernel is indeed a valid kernel. [Hint: The following fact may be useful. $||x-z||^2 = ||x||^2 - 2x^T z + ||z||^2$.]

4. **[18 points] One-class SVM**

   Given an unlabeled set of examples $\{x^{(1)}, \ldots, x^{(m)}\}$ the *one-class SVM algorithm* tries to find a direction w that maximally separates the data from the origin. [2]

   More precisely, it solves the (primal) optimization problem:

   $$\begin{array}{ll} \min_w & \frac{1}{2} w^\top w \\ \text{s.t.} & w^\top x^{(i)} \geq 1 \qquad \text{for all } i = 1, \ldots, m \end{array}$$

   A new test example x is labeled 1 if $w^\top x \geq 1$, and 0 otherwise.

   (a) **[9 points]** The primal optimization problem for the one-class SVM was given above. Write down the corresponding dual optimization problem. Simplify your answer as much as possible. In particular, w should not appear in your answer.

[1] Equivalently, $y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$, where the $\varepsilon^{(i)}$'s are distributed IID $\mathcal{N}(0, \sigma^2)$.

[2] This turns out to be useful for anomaly detection, but I assume you already have enough to keep you entertained for the 2h 15min of the midterm, and thus wouldn't want to read about it here. See the midterm solutions for details.

(b) [**4 points**] Can the one-class SVM be kernelized (both in training and testing)? Justify your answer.

(c) [**5 points**] Give an SMO-like algorithm to optimize the dual. I.e., give an algorithm that in every optimization step optimizes over the smallest possible subset of variables. Also give in closed-form the update equation for this subset of variables. You should also justify why it is sufficient to consider this many variables at a time in each step.

5. **[18 points] Uniform Convergence**

In this problem, we consider trying to estimate the mean of a biased coin toss. We will repeatedly toss the coin and keep a running estimate of the mean. We would like to prove that (with high probability), after some initial set of N tosses, the running estimate from that point on will *always* be accurate and never deviate too much from the true value.

More formally, let $X_i \sim \text{Bernoulli}(\phi)$ be IID random variables. Let $\hat{\phi}_n$ be our estimate for ϕ after n observations:

$$\hat{\phi}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

We'd like to show that after a certain number of coin flips, our estimates will stay close to the true value of ϕ. More formally, we'd like to show that for all $\gamma, \delta \in (0, 1/2]$, there exists a value N such that

$$P\left(\max_{n \geq N} |\phi - \hat{\phi}_n| > \gamma\right) \leq \delta.$$

Show that in order to make the guarantee above, it suffices to have $N = O(\frac{1}{\gamma^2}\log(\frac{1}{\delta\gamma}))$. You may need to use the fact that for $\gamma \in (0, 1/2]$, $\log(\frac{1}{1-\exp(-2\gamma^2)}) = O(\log(\frac{1}{\gamma}))$.

[Hint: Let A_n be the event that $|\phi - \hat{\phi}_n| > \gamma$ and consider taking a union bound over the set of events $A_n, A_{n+1}, A_{n+2}, \ldots$..]

6. **[40 points] Short Answers**

The following questions require a true/false accompanied by one sentence of explanation, or a reasonably short answer (usually at most 1-2 sentences or a figure).

To discourage random guessing, one point will be deducted for a wrong answer on multiple choice questions! Also, no credit will be given for answers without a correct explanation.

(a) [**5 points**] Let there be a binary classification problem with continuous-valued features. In Problem Set #1, you showed if we apply Gaussian discriminant analysis using the same covariance matrix Σ for both classes, then the resulting decision boundary will be linear. What will the decision boundary look like if we modeled the two classes using separate covariance matrices Σ_0 and Σ_1? (I.e., $x^{(i)}|y^{(i)} = b \sim \mathcal{N}(\mu_b, \Sigma_b)$, for $b = 0$ or 1.)

(b) [**5 points**] Consider a sequence of examples $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)})$. Assume that for all i we have $\|x^{(i)}\| \leq D$ and that the data are linearly separated with a margin γ. Suppose that the perceptron algorithm makes exactly $(D/\gamma)^2$ mistakes on this sequence of examples. Now, suppose we use a feature mapping $\phi(\cdot)$ to a higher dimensional space and use the corresponding kernel perceptron algorithm on the same sequence of data (now in the higher-dimensional feature space). Then the kernel perceptron (implicitly operating in this higher dimensional feature space) will make a number of mistakes that is

i. strictly less than $(D/\gamma)^2$.
ii. equal to $(D/\gamma)^2$.
iii. strictly more than $(D/\gamma)^2$.
iv. impossible to say from the given information.

(c) [**5 points**] Let any $x^{(1)}, x^{(2)}, x^{(3)} \in \mathbb{R}^p$ be given $(x^{(1)} \neq x^{(2)}, x^{(1)} \neq x^{(3)}, x^{(2)} \neq x^{(3)})$. Also let any $z^{(1)}, z^{(2)}, z^{(3)} \in \mathbb{R}^q$ be fixed. Then there exists a valid Mercer kernel $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ such that for all $i, j \in \{1, 2, 3\}$ we have $K(x^{(i)}, x^{(j)}) = (z^{(i)})^\top z^{(j)}$. True or False?

(d) [**5 points**] Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be defined according to $f(x) = \frac{1}{2}x^\top Ax + b^\top x + c$, where A is symmetric positive definite. Suppose we use Newton's method to minimize f. Show that Newton's method will find the optimum in exactly one iteration. You may assume that Newton's method is initialized with $\vec{0}$.

(e) [**5 points**] Consider binary classification, and let the input domain be $\mathcal{X} = \{0, 1\}^n$, i.e., the space of all n-dimensional bit vectors. Thus, each sample x has n binary-valued features. Let $\mathcal{H}_n$ be the class of all boolean functions over the input space. What is $|\mathcal{H}_n|$ and $VC(\mathcal{H}_n)$?

(f) [**5 points**] Suppose an ℓ_1-regularized SVM (with regularization parameter $C > 0$) is trained on a dataset that is linearly separable. Because the data is linearly separable, to minimize the primal objective, the SVM algorithm will set all the slack variables to zero. Thus, the weight vector w obtained will be the same no matter what regularization parameter C is used (so long as it is strictly bigger than zero). True or false?

(g) [**5 points**] Consider using hold-out cross validation (using 70% of the data for training, 30% for hold-out CV) to select the bandwidth parameter τ for locally weighted linear regression. As the number of training examples m increases, would you expect the value of τ selected by the algorithm to generally become larger, smaller, or neither of the above? For this problem, assume that (the expected value of) y is a non-linear function of x.

(h) [**5 points**] Consider a feature selection problem in which the mutual information $MI(x_i, y) = 0$ for all features x_i. Also for every subset of features $S_i = \{x_{i_1}, \cdots, x_{i_k}\}$ of size $< n/2$ we have $MI(S_i, y) = 0$.[3] However there is a subset S^* of size exactly $n/2$ such that $MI(S^*, y) = 1$. I.e. this subset of features allows us to predict y correctly. Of the three feature selection algorithms listed below, which one do you expect to work best on this dataset?

i. Forward Search.
ii. Backward Search.
iii. Filtering using mutual information $MI(x_i, y)$.
iv. All three are expected to perform reasonably well.

[3] $MI(S_i, y) = \sum_{S_i} \sum_y P(S_i, y) \log(P(S_i, y)/P(S_i)P(y))$, where the first summation is over all possible values of the features in S_i.