Pratice Set - Minor 1 Exam

January 26, 2017

1. Consider learning a logistic regression model over a set of data points $\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$. Recall that the log-likelihood for logistic regression is given by:

$$LL(\theta) = \sum_{i=1}^{m} \Big[y^{(i)} log(h_\theta(x^{(i)})) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)})) \Big] \tag{1}$$

where $h_\theta(x^{(i)}) = g(\theta^T x^{(i)})$ and $g(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function. Show that $LL(\theta)$ is a concave function of θ. You should carry out your calculations from first principles (not directly use any results derived in class). You can use the fact that $\frac{\partial g(z)}{\partial z} = g(z)(1 - g(z))$.

Hint: Try showing that the associated Hessian matrix is negative semi-definite. A $n \times n$ matrix H is negative semi-definite iff $\forall z \in \mathcal{R}^n$, $z^T H z \leq 0$. Also, recall that $z^T H z = \sum_{j,k} z_j H_{jk} z_k$.

2. Overfitting refers to the phenomenon of an algorithm overtly fitting the patterns in the training data which may not generalize well. On the other hand, underfitting refers to the phenomenon of an algorithm not being able to capture even the desriable patterns existing in the data (due to limitations on its representational power). In the description below, training data is the data over which the model is learned, and test data is some unseen data coming from the same disribution. Consider learning three different classifiers C_1, C_2, C_3 on a given data set such that C_1 has high training as well as test accuracies, C_2 has high training accuracy but low test accuracy, whereas C_3 has low training as well as test accuracies. Which one of the following statements is correct?

   (a) C_1 is overfitting whereas C_2 is underfitting.

   (b) C_1 is overfitting whereas C_3 is underfitting.

   (c) C_2 is overfitting whereas C_3 is underfitting.

   (d) C_2 is underfitting whereas C_3 is overfitting.

   (e) None of the above.

   Justify your answer.

3. Generalized Linear Models (GLMs) assume that the target variable y (conditioned on x) is distributed according to a member of the exponential family:

$$p(y; \eta) = b(y) exp(\eta y - a(\eta)), \tag{2}$$

where $\eta = \theta^T x$. Further, $h_\theta(x) = E[y|x; \eta]$. Other symbols are as defined in the class.

Show that linear regression under the assumption of i.i.d. Gaussian noise belongs to the class of GLMs. You can assume that noise (ϵ) is distributed as $\epsilon \sim \mathcal{N}(\mu, 1)$. What are the values of $b(y)$ and $a(\eta)$? Also, describe the relationship between μ and η.

4. **You should try to solve this question after the class on Friday the 27th**.
Suppose you are minimizing a convex objective function using gradient descent and the algorithm has not converged even after 10,000 steps. What might be the possible reasons? Is there any alternate algorithm (other than gradient descent) that may lead to faster convergence? Argue.

5. **You should try to solve this question after the class on Friday the 27th**.
Given a set of data points $\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$ with real valued target variable, recall that linear regression tries to optimize the error function $J(\theta)$ given as:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2$$

Show that error function can be written as $J(\theta) = \frac{1}{2}(X\theta - Y)^T (X\theta - Y)$ where $X \in \mathcal{R}^{m \times (n+1)}$, $\theta \in \mathcal{R}^{(n+1)}$ and $Y \in \mathcal{R}^m$ are suitably defined matrices. Using this expression, derive the analytical solution for θ^* which minimizes $J(\theta)$ from first principles using matrix calculus. You are not allowed to make use of the trace operator. You can use the fact that $\Delta_\theta(\theta^T A \theta) = 2A\theta$, where A is some $(n+1) \times (n+1)$ matrix.

6. **You should be in a position to solve this question after the class on Saturday the 28th.**
Consider a learning task where $P(x|y=0)$ and $P(x|y=1)$ $(x \in \{0\} \cup \mathbb{Z}_+)$ are modeled as Poisson random variables with mean parameters λ_0 and λ_1, respectively. Note that the learning problem is described using a single dimensional feature vector (scalar) x. Show that $P(y=1|x)$ takes the form of a logistic function i.e. $P(y=1|x) = \frac{1}{1+e^{-\theta^T x}}$. You should derive an expression for θ in terms of λ_0 and λ_1. Do not forget to include the intercept (θ_0) term.

   Note: Poisson distribution is a discrete distribution (parameterized by mean parameter λ) defined as $P(x=k) = \frac{e^{-\lambda}\lambda^k}{k!}(k \geq 0)$.

7. **You should be in a position to solve this question after the class on Saturday the 28th.**
Consider a binary classification problem with continuous-valued features. We showed in class that if we apply Gaussian discriminant analysis using the same covariance matrix Σ for both the classes, then the resulting decision boundary will be linear. Consider the case when two class distributions are modeled using separate covariance matrices, Σ_0 and Σ_1, i.e. $(x^{(i)}|y^{(i)}=0) \sim \mathcal{N}(\mu_0, \Sigma_0)$ and $(x^{(i)}|y^{(i)}=1) \sim \mathcal{N}(\mu_1, \Sigma_1)$, where $\Sigma_0 \neq \Sigma_1$. We will assume that Σ_0 and Σ_1 are diagonal matrices with non-zero diagonal entries. Further, diagonal entries of Σ_0 are less than equal to the respective diagonal entries of Σ_1 i.e.,

$$\begin{aligned} (\Sigma_0)_{jk} &= (\Sigma_1)_{jk} = 0 \quad for \quad j \neq k \\ (\Sigma_0)_{jk} &\leq (\Sigma_1)_{jk} \quad for \quad j = k \end{aligned}$$

   Describe the form of the decision boundary under the above assumptions. Specifically, state whether the decision boundary contains any quadratic terms, and if yes, characterize them as precisely as possible. Mathematically justify your answer. Describe the shape of the boundary (as precisely as possible) when there are only two features. Justify your answer. You can make use of the following:

   - If A is a diagonal matrix with i^{th} entry given by a_{ii}, A^{-1} is also diagonal with i^{th} entry given by $\frac{1}{a_{ii}}$.
   - $x^T A x = \sum_{jk} x_j A_{jk} x_k$, where A is an $n \times n$ matrix, A_{jk} is the jk^{th} entry of A and $x \in \mathcal{R}^n$.
   - If $x \sim \mathcal{N}(\mu, \Sigma)$, then $P(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp(\frac{-1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))$