COL 774: Machine Learning
Minor 1 Examination

Thursday February 2, 2017

Notes:

- **Time: 2:25 pm to 3:35 pm. Total Questions: 5. Maximum Points: 20.**
- **This question paper has printed material on both sides.**
- **This exam is closed book/notes.**
- **Some questions may be harder than others. Use your time wisely.**
- **Start each answer on a new page.**
- **You need justify all your answers. Answers without justification may not get full points.**
- **In the following problems, m will denote the number of examples and n the number of features, unless otherwise stated.**

1. In the class, we formulated logistic regression as a two class classifcation problem. In particular, each label $y^{(i)}$ belonged to the set $\{0, 1\}$. Given the parameter vector $\theta \in \mathcal{R}^{n+1}$, the probability of $y|x$ was defined using the Bernoulli distribution. Specifically, the distribution took the following form:
$P(y = 1|x; \theta) = \frac{1}{1+e^{-\theta^T x}}$
In this problem, we will extend the above model to a multi-class setting. Assume that labels set is given by $\{1, 2, \cdots, r\}$ (i.e., consisting of r different labels). The problem is now characterized by a set of parameter vectors $\Theta = \{\theta_{(k)}\}_{k=1}^{r}$ where each $\theta_{(k)} \in \mathcal{R}^{n+1}$. The natural extension of the Bernoulli distribution to the multi-class setting would be to a multinomial distribution. So, the probability distribution of $(y = k|x)$ is defined as:

$$P(y = k|x; \Theta) = \frac{1}{Z} * e^{(\theta_{(k)})^T x} \tag{1}$$

Here, $Z = \sum_k e^{(\theta_{(k)})^T x}$ is the normalization constant. This is called the softmax regression model.

   - Write the expression for log-likelihood $LL(\Theta)$ for the softmax regression model as described above. You should write it in a form which enables gradient computation.
   - Find the expression for the gradient $\nabla_{\theta_{(k)}} LL(\Theta)$ for each of the parameter vectors $\theta_{(k)}$. In particular, show that the gradient with respect to the $\theta_{(k)}$ parameters can be written as $\sum_{i=1}^{m}(\mathbb{1}\{y^{(i)} = k\} - h_{\theta_{(k)}}(x^{(i)}))x^{(i)}$ for suitably defined $h_{\theta_{(k)}}(x^{(i)})$ functions. Here, $\mathbb{1}$ denotes the indicator function.

2. Let Σ denote the covariance matrix for a set of n random variables. Recall that Σ is a symmetric matrix. We say that a symmetric matrix $A \in \mathcal{R}^{n \times n}$ is positive semi-definite, if $\forall z \in \mathcal{R}^n$, we have $z^T A z \geq 0$. Show that the covariance matrix Σ is positive semi-definite. You may find it helpful to use the linearity of expectations, i.e. if $Y_1, Y_2, \cdots, Y_r$ are r random variables, then $E[\sum_{k=1}^{r} \lambda_k Y_k] = \sum_{k=1}^{r} \lambda_k E[Y_k]$, where λ_k's are constants.

3. Recall that linear regression tries to optimize the error function defined as $J(\theta) = \frac{1}{2m}\sum_i (y^{(i)} - \theta^T x^{(i)})^2$ where the symbols are as defined in class. Consider defining an altnerate objective function $J_{reg}(\theta) = \frac{1}{2m}\big(\sum_i (y^{(i)} - \theta^T x^{(i)})^2 + \lambda \theta^T \theta\big)$ where λ is a constant. Re-express $J_{reg}(\theta)$ in terms of suitably defined matrices X, Y, θ where $X \in \mathcal{R}^{m \times (n+1)}, Y \in \mathcal{R}^m, \theta \in \mathcal{R}^{(n+1)}$. Clearly specify what the matrices are in your expression. Now, use your expression to analytically compute the value of the θ parameters which minimize $J_{reg}(\theta)$.
Note: This is called the regularized formulation of least squares regression and is often used to combat overfitting. We will discuss more about this at a later point during the course.

4. Consider a binary classification problem with continuous-valued features. Consider applying GDA to this problem. Let the parameters of the model be given as $\Theta = (\phi, \mu_0, \mu_1, \Sigma_0, \Sigma_1)$, where the symbols are as defined in class. Let us make the additional assumption that the two covariance matrices are identical i.e., $\Sigma_0 = \Sigma_1 = \Sigma$. Under this assumption, we (almost) showed in class that the decision boundary enforced by GDA will be a linear function of x. In other words, the decision boundary can be written as $\theta^T x = 0$ where $\theta \in \mathcal{R}^{(n+1)}$. Explicitly derive the value of the θ parameters (including the intercept term) in terms of the GDA parameters $\phi, \mu_0, \mu_1, \Sigma$.
Recall: If $x \sim \mathcal{N}(\mu, \Sigma)$, then $P(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp(\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$

5. Suppose that you are working with a supervised machine learning (classification) problem and you have access to m examples $\{x^{(i)}, y^{(i)}\}_{i=1}^m$ coming from some (unknown) underlying distribution. Assume that these are the only examples that you have access to. You would like to build a model to predict the y's from x's so as to generalize well to future (unseen) data. You have a bunch of learning algorithms at hand and you would like to find out which of these is most suited for this problem. Since, you don't know yet about much about machine learning, the only thing that you can do is train each algorithm on a subset of the available data (called the training data), predict the accuracy of the learned model on a (possibly different) subset of the available data (called the test data) and finally declare as algorithm of choice the one having the highest accuracy on the test data. Now, consider the following scenarios to decide the train/test subsets:

   (a) You choose the entire set of examples for training as well as for testing.

   (b) You randomly choose half of the examples as the training set and then independently (at random) choose another half as the test set. The two sets could be overlapping since they are chosen independent of each other.

   (c) You decide a number $m' \leq m$. You randomly choose m' examples as your training set. The remaining $m - m'$ examples become your test set. The two sets of are disjoint in this case.

   Answer the following:

   - Which of the above scenarios is likely to give you best performance on the unseen data? Argue for each case.
   - If one decides to go for alternative (c), describe the considerations that you will make in choosing m'. Note that choosing m' decides the size of the training as well as the test set (since the total number of examples is fixed).