COL 774: Machine Learning
Minor 1 Examination

Monday February 5, 2018

Notes:

- **Time: 2:25 pm to 3:35 pm. Total Questions: 5. Maximum Points: 20.**
- **This exam is closed book/notes.**
- **This question paper has printed material on both sides.**
- **Each question carries 5 points.**
- **Some questions may be harder than others. Use your time wisely.**
- **Start each answer on a new page.**
- **Justify all your answers. Answers without justification may not get full points.**
- **We will use the notation as described in class unless otherwise stated.**

1. Consider optimizing a function $f(\theta)$ which is a quadratic function of its parameters. Show that Newton's method for optimizing $f(\theta)$ will converge in a single iteration. Recall that the update rule for Newton's method is given as:

$$\theta^{(t+1)} = \theta^{(t)} - H^{-1}\nabla_\theta f(\theta^{(t)}) \tag{1}$$

where H denotes the Hessian matrix. Other symbols are as defined in class. You can assume that the matrix H is invertible. You should prove it for the general case when $\theta \in \mathcal{R}^n$. If you can't prove it for the general case, see if you can prove it when $\theta \in \mathcal{R}$ (to get some partial points).

2. Recall that locally weighted linear regression optimizes the following error metric:

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

where $w^{(i)}$'s are as defined in class. m denotes the number of examples. Show that the error function $J(\theta)$ as defined above is convex in θ. You can use the fact that a function is convex if the corresponding Hessian matrix H is positive semi-definite, i.e. $z^T H z \geq 0, \forall z \in \mathcal{R}^n$, n being the dimension of the parameters.

3. Consider minimizing a function $f(\theta)$ using Stochastic Gradient Descent (SGD). Let m denote the number of examples and n denote the number of attributes. Assume that you are using a single example in each iteration of SGD for parameter update.

   (a) Write down the pseudocode for stochastic gradient descent in the above setting.

   (b) Clearly describe (using pseudocode) how will you detect the convergence of your algorithm. Note that a single iteration of SGD may not give enough information for detecting convergence since it only looks at one example. Your convergence detection code should be efficient i.e., its running time should not depend on m (number of examples).

4. Consider modeling a binary classification problem using GDA. 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. We will make the further assumption of $\Sigma_1 = \Sigma_0 = \Sigma$.

   (a) Show that in the above setting, the probability distribution learned by GDA takes the form of the logistic function, i.e.,

$$P(y=1|x;\Theta) = \frac{1}{1+e^{-\theta^T x}} \tag{3}$$

   where θ can be expressed as a function of the Θ parameters. You do not have to derive the exact functional form for θ but only prove (albeit mathematically) that the distribution can be written in the above form.

(b) Above result shows that under the assumption of identical co-variance matrices for GDA, the distribution of $y|x$ specified by GDA and logistic regression has the same form. In this setting, which of the two models, GDA or logistic regression, do you think makes stronger assumptions about how the underlying data is distirbuted. Under what circumstances would you expect logistic regression to perform better than GDA? Why?

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

$$p(y;\eta) = b(y)exp(\eta y - a(\eta)), \tag{4}$$

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

(a) Show that the Bernoulli distribution with parameter ϕ belongs to the exponential family. Clearly describe the relationship between the parameters ϕ and η. Also, find out the expressions for $b(y)$ and $a(\eta)$.

(b) Use the above result to show that logistic regression belongs to the class of GLMs.