COL 774: Machine Learning
Minor 1 Examination

Monday February 6, 2019

Notes:

- **Time: 3:55 pm to 5:05 pm. Total Questions: 5. Maximum Points: 25.**
- **This exam is closed book/notes. One A4 sized sheet of handwritten notes is allowed as annouced on class mailing list.**
- **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 a learning problem with m training examples given as $\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$. Assume that we are trying to learn a logistic regression model. Recall that under the logistic regression model, the log-likelihood $LL(\theta)$ can be written as:

$$LL(\theta) = \sum_{i=1}^{m} log(P(y^{(i)}|x^{(i)};\theta)) \tag{1}$$

   where $P(y^{(i)} = 1|x^{(i)};\theta)$ is given by the sigmoid function $\frac{1}{1+e^{-\theta^T x^{(i)}}}$. Show that $LL(\theta)$ is a concave function of the θ parameters. You should prove it from first principles and not directly use any gradient expressions derived in class. Hint: You can show that the corresponding Hessian matrix is negative semi-definite. Feel free to use the fact that Hessian matrix is symmetric in this case (and in general). Recall that a symmetric matrix $A \in \mathcal{R}^{n \times n}$ is negative semi-definite if $z^t A z \leq 0$, $\forall z \in \mathcal{R}^n$.

2. Consider learning a locally weighted linear regression model in one dimensional setting, i.e., $x, y \in \mathcal{R}$. Let the set of training examples be denoted by T_r. Assume that the total number of examples in the training set is given by 100, i.e., $|T_r| = 100$. Describe the process of validation to estimate the best value of τ parameter in the model. Now, assume that you are also given a separate test T_e. Report the error over the test set using your best value of τ. You should write your procedure in the form of detailed pseudo-code so that one can easily translate this into a piece of actual code. You can assume following are available to you:

   - A sub-routine $M = learnLWR(S, x, \tau)$ which returns the locally weighted regression model M learned using the subset of examples $S \subseteq T_r$ at point x and using the parameter value τ.
   - A sub-routine $\hat{y} = M(x)$ which returns $\hat{y}$ as the prediction of model M (as learnt above) at point x.
   - A sub-routine $\epsilon = Error(\hat{y}, y)$ which returns the error of prediction ϵ based on the predicted value $\hat{y}$ and the target value y.

   Recall that the weights $w^{(i)}$'s in locally weighted linear regression are chosen as $w^{(i)} = \exp(\frac{-1}{2\tau^2}(x - x^{(i)})^2)$ where τ is the bandwidth parameter.

3. Consider a learning problem over m training examples given as $\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$. Let the training set be denoted as T_r. Assume a generative model where we are generating both $x^{(i)}$ and $y^{(i)}$'s under some modelling assumptions. Assume that we are learning the parameters of the model by maximizing the log-likelihood, denoted by $LL(\theta)$ where θ denotes the set of parameters in the model. Further, assume that we are using the following version of stochastic gradient ascent as our learning algorithm:

(a) Randomly sample an example $e^{(i)}$, denoted as $(x^{(i)}, y^{(i)})$ from the training set T_r.

(b) Let the gradient computed using example $e^{(i)}$ be given as $g^{(i)}$ (at the current set of parameters).

(c) Update the parameters using the equation $\theta^{(t+1)} = \theta^{(t)} + \eta g^{(i)}$. Here, $\theta^{(t)}$ denotes the set of parameters at iteration t and η is the learning rate.

The three steps above define one iteration of the learning algorithm. We repeat these steps until convergence.

Now, let g denote the value of the gradient (at the current set of parameters) computed using all the examples. Show that $\forall j, g_j = m \times E[g_j^{(i)}]$ where the expectation is taken over the uniform distribution defined over the set of all training examples. Here, g_j and $g_j^{(i)}$ denote the partial derivatives corresponding to the j^{th} components in g and $g^{(i)}$, respectively. Does your proof need to make the assumption about training examples being independent of each other? If yes, precisely indicate where you made that assumption. What would happen if you did not make this assumption? Note that this proof is at the heart of convergence of SGD to close vicinity of a local optima.

4. Consider modeling a binary classification problem using Gaussian Discriminant Analysis (GDA). Let $x \in \mathcal{R}^n$. 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. Assume that both Σ_0 and Σ_1 are diagonal matrices.

   (a) Show that the shape of the decision boundary is quadratic in the above model. Derive the coefficients of the quadratic terms in the equation of the decision boundary in terms of diagonal entries of Σ_1 and Σ_0, written as $(\Sigma_1)_{jj}$ and $(\Sigma_0)_{jj}$, respectively (j denotes the index of the diagonal element). You should simplify the form of your coefficients as much as possible.

   (b) Now assume that the co-variance matrices are tied with each other, i.e., $\Sigma_0 = \Sigma_1$. Show that decision boundary is linear in this case.

   (c) In this part, assume that you are solving the above problem such that $x \in \mathcal{R}^2$. Then, give a crisp name to the shape of your decision boundary in each of the parts above (your name should be as precise as possible).

   Recall that for $x \in \mathcal{R}^n$, if $x \sim N(\mu, \Sigma)$, then $P(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp(\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$ where symbols are as defined in class.

5. Consider a learning problem where the target value given input features is distributed using Poisson distribution, i.e., $y|x \sim Poisson(\lambda)$. Assume that the learning model has some underlying set of parameters θ and the relationship between λ, θ and x is given as $\lambda = \exp(\theta^T x)$. Given m training examples, $\{x^{(i)}, y^{(i)}\}_{i=1}^m$, consider learning the parameters of this model using gradient ascent. Show that the parameter update rule can be written as:

$$\theta^{(t+1)} = \theta^{(t)} + \eta \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)}))x^{(i)} \tag{2}$$

   here $\theta^{(t)}$ represents the set of parameters at iteration t, η is the learning rate and other symbols are as defined in class. What is the value of $h_\theta(x^{(i)})$ (in terms of θ parameters) which satisfies the above equation (and also makes intuitive sense)? Show that $h_\theta(x^{(i)}) = E[y^{(i)}|x^{(i)}; \theta]$. Note this is another model which falls in the class of GLMs. Recall that Poisson distribution is a discrete distribution parameterized by mean parameter λ such that if $y \sim Poisson(\lambda)$, then $P(y = k) = \frac{e^{-\lambda}\lambda^k}{k!} (k \geq 0)$.