STAT 6910-003 – SLDM II – Homework #3

Due: 5:00 PM 11/2/18

1. **Maximum Likelihood Estimation (14 pts)**

   (a) Let $X_1, \ldots, X_n$ be i.i.d. sample from a Poisson distribution with parameter λ, i.e.

   $$P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

   i. (2 pts) Write down the likelihood function $L(\lambda)$.

   ii. (2 pts) Write down the log-likelihood function $\ell(\lambda)$.

   iii. (3 pts) Find the maximum likelihood estimate (MLE) of the parameter λ.

   (b) (7 pts) Let $X_1, \ldots, X_n$ be an i.i.d. sample from an exponential distribution with the density function

   $$p(x; \beta) = \frac{1}{\beta} e^{-\frac{x}{\beta}},\ 0 \leq x < \infty.$$

   Find the MLE of the parameter β. Given what you know about the role that β plays in the exponential distribution, does the MLE make sense? Why or why not?

2. **Logistic Regression as ERM (5 pts).** Consider training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ for binary classification and assume $y_i \in \{-1, 1\}$. Show that if $L(y, t) = \log(1 + \exp(-yt))$,then

   $$\frac{1}{n} \sum_{i=1}^{n} L(y_i, \boldsymbol{w}^T \boldsymbol{x}_i + b)$$

   is proportional to the negative log-likelihood for logistic regression. Therefore ERM with the logistic loss is equivalent to the maximum likelihood approach to logistic regression.
   Clarification: In the above expression, y is assumed to be -1 or 1. In the notes, we had $y \in \{0, 1\}$. So all you need to do is rewrite the negative log-likelihood for logistic regression using the ± 1 label convention and simplify that formula until it looks like the formula above.

3. **Convexity and Optimization (24 pts).** Let $f : \mathbb{R}^d \to \mathbb{R}$.

   (a) (7 pts) Show that if f is strictly convex, then f has at most one global minimizer.

   (b) (7 pts) Use the Hessian to give a simple proof that the sum of two convex functions is convex. You may assume that the two functions are twice continuously differentiable.

   (c) (7 pts) Consider the function $f(\boldsymbol{x}) = \frac{1}{2} \boldsymbol{x}^T A \boldsymbol{x} + \boldsymbol{b}^T \boldsymbol{x} + c$ where A is a symmetric $d \times d$ matrix. Derive the Hessian of f. Under what conditions on A is f convex? Strictly convex?

   (d) (3 pts) Let $J(\boldsymbol{\theta})$ be a twice continuously differentiable function. Derive the update step for Newton's method from the second order approximation of $J(\boldsymbol{\theta})$ (see lecture slides for equations for both the update step and the second order approximation).

4. **Logistic regression Hessian (15 pts).** Determine a formula for the gradient and the Hessian of the regularized logistic regression objective function. Argue that the objective function

   $$J(\boldsymbol{\theta}) = -\ell(\boldsymbol{\theta}) + \lambda ||\boldsymbol{\theta}||^2$$

   is convex when $\lambda \geq 0$, and that for $\lambda > 0$, the objective function is strictly convex.
   Hints: The following conventions and properties regarding vector differentiation may be useful. The properties can be easily verified from definitions. Try to avoid long, tedious calculations.

- If $f : \mathbb{R}^n \to \mathbb{R}$, then we adopt the convention

$$\frac{\partial f(\boldsymbol{z})}{\partial \boldsymbol{z}} := \nabla f(\boldsymbol{z}).$$

- If $f : \mathbb{R}^n \to \mathbb{R}^m$, adopt the convention

$$\frac{\partial f(\boldsymbol{z})}{\partial \boldsymbol{z}^T} := \left(\frac{\partial f(\boldsymbol{z})}{\partial \boldsymbol{z}}\right)^T.$$

- Given these conventions, it follows that the Hessian H of J is

$$H = \frac{\partial}{\partial \boldsymbol{\theta}^T}\left(\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right),$$

which is often denoted more concisely as

$$\frac{\partial^2 J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}.$$

- (One form of a multivariate chain rule): If $f(\boldsymbol{z}) = g(h(\boldsymbol{z}))$ where $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}$, then

$$\nabla f(\boldsymbol{z}) = \nabla h(\boldsymbol{z}) \cdot g'(h(\boldsymbol{z})).$$

5. **ERM and Stochastic Gradient Descent (10 pts).** Given training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, define the empirical risk for either a regression or classification problem as

$$\hat{R}(f_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i.f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right).$$

Write pseudocode describing how you would implement stochastic gradient descent to minimize $\hat{R}(f_{\boldsymbol{\theta}})$ with respect to θ. Assume a fixed mini-batch size of m and assume that the step size α is fixed for each epoch.

6. **Handwritten digit classification with logistic regression (22 pts).** Download the file `mnist_49_3000.mat` from the Homework 3 assignment. This is a Matlab data file that contains a subset of the MNIST handwritten digit dataset, which is a well-known benchmark dataset for classification. This subset contains examples of the digits 4 and 9.
The data file contains variables $\boldsymbol{x}$ and y, with the former containing the image of the digit (reshaped into column vector form) and the latter containing the corresponding label ($y \in \{-1, 1\}$). To visualize an image, you will need to reshape the column vector into a square image. You should be able to find methods for loading the data file and for reshaping the vector in your preferred language through a Google search. If you're struggling to find something that works, you may ask for suggestions on Piazza.
Implement Newton's method to find a minimizer of the regularized negative log likelihood for logistic regression: $J(\boldsymbol{\theta}) = -\ell(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|^2$. Try setting $\lambda = 10$. Use the first 2000 examples as training data and the last 1000 as test data.

   (a) (6 pts) Report the test error, your termination criterion (you may choose), how you initialized $\boldsymbol{\theta}_0$, and the value of the objective function at the optimum.

   (b) (10 pts) Generate a figure displaying 20 images in a 4×5 array. These images should be the 20 misclassified images for which the logistic regression classifier was most confident about its prediction. You will have to define a notion of confidence in a reasonable way and explain how you define it. In the title of each subplot, indicate the true label of the image. What you should expect to see is a bunch of 4s that look kind of like 9s and vice versa.

   (c) (6 pts) Include your well-organized, clearly commented code.

7. How long did this assignment take you? (5 pts)

8. Type up homework solutions (5 pts)