Homework 1

Due Date: March 19, 2017

Problem 1. For parameter w, try to prove that logistic regression function $y = \frac{1}{1+e^{-(\boldsymbol{w^T x}+b)}}$ is non-convex, but its logarithmic likelihood function $l(\boldsymbol{w}) = \sum_{i=1}^{m}(-y_i(\boldsymbol{w^T x_i} + b) + \ln(1 + e^{\boldsymbol{w^T x_i}+b}))$ is convex.

Problem 2. Using the technique of Lagrange multipliers, show that minimization of the regularized error function

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w^T}\boldsymbol{\phi}(x_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

is equivalent to minimizing the unregularized sum-of-squares error

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w^T}\boldsymbol{\phi}(x_n)\}^2$$

subject to the constraint $\sum_{j=1}^{M}|w_j|^q \leq \eta$.Discuss the relationship between the parameters η and λ.

Problem 3. Consider a data set in which each data point t_n is associated with a weighting factor $r_n > 0$, so that the sum-of-squares error function becomes

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2.$$

where $\phi(\mathbf{x}_n)$ is basis function. Find an expression for the solution $\mathbf{w}^*$ that minimizes this error function.

Problem 4. Multinomial logistic regression is a classification method that generalizes logistic regression to multiclass problems. It has the form

$$p(y = c|\mathbf{x}, \mathbf{W}) = \frac{exp(w_{c0} + \mathbf{w_c}^T\mathbf{x})}{\sum_{k=1}^{C} exp(w_{k0} + \mathbf{w_k}^T\mathbf{x})},$$

where C is the number of classes, and $\mathbf{W}$ is a C x (d+1) weight matrix, and d is the dimension of input vector $\mathbf{x}$. Suppose we are given a set of training data $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$, and we want to learn a set of weight vectors that maximize the conditional likelihood of the output labels $\{y_i\}_{i=1}^n$, given the input data $\{\mathbf{x}_i\}_{i=1}^n$ and $\mathbf{W}$. That is, we want to solve the following optimization problem (assuming the data points are i.i.d).

$$\mathbf{W}* = \underset{\mathbf{W}}{argmax} \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{W})$$

(a) Derive the conditional log-likelihood function for the multinomial logistic regression model. You may denote this function as l(W)

(b) Derive the gradient of l(W) with respect to the weight vector of class c ($\mathbf{w}_c$). That is, derive $\nabla_{\mathbf{w}_c} l(W)$. You may denote this function as gradient $g_c(W)$. Note: The gradient of a function $f(\mathbf{x})$ with respect to a vector $\mathbf{x}$ is also a vector, whose i-th entry is defined as $\frac{\partial^2 f(\mathbf{x})}{\partial x_i}$.

(c) Derive the block submatrix of the Hessian with respect to weight vector of class c ($\mathbf{w}_c$) and class c' ($\mathbf{w}_{c'}$). You may denote this function as $H_{c,c'}(\mathbf{W})$. Note: The Hessian of a function $f(\mathbf{x})$ with respect to vector $\mathbf{x}$ is a matrix, whose {i, j}th entry is defined as $\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}$. In this case, we are asking a block submatrix of Hessian of the conditional log-likelihood function, taken with respect to only two classes c and c'. The {i,j}th entry of the submatrix is defined as $\frac{\partial^2 l(\mathbf{W})}{\partial w_{ci} \partial w_{c'j}}$.

Problem 5. *This is a programming assignment.* You are required to use linear regression to find the relation between **price** and **sqft_living** in **train.csv**, and predict the house price based on **sqft_living** in **test.csv**. Please implement the following three different methods to find the best fitting function:

1. Gradient Descent
2. Newton's method
3. Normal Equation

After fitting a line to the training data,

1. You are required to make a scatterplot of **price** vs **sqft_living**, and plot your fitting line on this scatterplot.
2. You are required to use your fitting line to predict the house price based on **sqft_living** in **test.csv**, and compute the RMSE(root-mean-square error) of your prediction.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \left(y_i^{predict} - y_i^{true}\right)^2}{N}} \tag{1}$$

, where N is the number of test samples, $y_i^{predict}$ is the prediction price for the i^{th} sample, y_i^{true} is the true price for the i^{th} sample.

In your report, you are required to compare the three methods with the help of three **plots**, three **fitting functions** and three **RMSEs**.