CS 229, Public Course
Problem Set #1: Supervised Learning

---

1. **Newton's method for computing least squares**

   In this problem, we will prove that if we use Newton's method solve the least squares optimization problem, then we only need one iteration to converge to θ^*.

   (a) Find the Hessian of the cost function $J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})^2$.

   (b) Show that the first iteration of Newton's method gives us $\theta^\star = (X^T X)^{-1} X^T \vec{y}$, the solution to our least squares problem.

2. **Locally-weighted logistic regression**

   In this problem you will implement a locally-weighted version of logistic regression, where we weight different training examples differently according to the query point. The locally-weighted logistic regression problem is to maximize

$$\ell(\theta) = -\frac{\lambda}{2}\theta^T\theta + \sum_{i=1}^{m} w^{(i)}\left[y^{(i)}\log h_\theta(x^{(i)}) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right].$$

   The $-\frac{\lambda}{2}\theta^T\theta$ here is what is known as a regularization parameter, which will be discussed in a future lecture, but which we include here because it is needed for Newton's method to perform well on this task. For the entirety of this problem you can use the value $\lambda = 0.0001$.

   Using this definition, the gradient of $\ell(\theta)$ is given by

$$\nabla_\theta \ell(\theta) = X^T z - \lambda\theta$$

   where $z \in \mathbb{R}^m$ is defined by

$$z_i = w^{(i)}(y^{(i)} - h_\theta(x^{(i)}))$$

   and the Hessian is given by

$$H = X^T D X - \lambda I$$

   where $D \in \mathbb{R}^{m\times m}$ is a diagonal matrix with

$$D_{ii} = -w^{(i)} h_\theta(x^{(i)})(1 - h_\theta(x^{(i)}))$$

   For the sake of this problem you can just use the above formulas, but you should try to derive these results for yourself as well.

   Given a query point x, we choose compute the weights

$$w^{(i)} = \exp\left(-\frac{||x - x^{(i)}||^2}{2\tau^2}\right).$$

   Much like the locally weighted linear regression that was discussed in class, this weighting scheme gives more when the "nearby" points when predicting the class of a new example.

(a) Implement the Newton-Raphson algorithm for optimizing $\ell(\theta)$ for a new query point x, and use this to predict the class of x.

The `q2/` directory contains data and code for this problem. You should implement the `y = lwlr(X_train, y_train, x, tau)` function in the `lwlr.m` file. This function takes as input the training set (the `X_train` and `y_train` matrices, in the form described in the class notes), a new query point `x` and the weight bandwitdh `tau`. Given this input the function should 1) compute weights $w^{(i)}$ for each training example, using the formula above, 2) maximize $\ell(\theta)$ using Newton's method, and finally 3) output $y = 1\{h_\theta(x) > 0.5\}$ as the prediction.

We provide two additional functions that might help. The `[X_train, y_train] = load_data;` function will load the matrices from files in the `data/` folder. The function `plot_lwlr(X_train, y_train, tau, resolution)` will plot the resulting classifier (assuming you have properly implemented `lwlr.m`). This function evaluates the locally weighted logistic regression classifier over a large grid of points and plots the resulting prediction as blue (predicting $y = 0$) or red (predicting $y = 1$). Depending on how fast your `lwlr` function is, creating the plot might take some time, so we recommend debugging your code with `resolution = 50;` and later increase it to at least 200 to get a better idea of the decision boundary.

(b) Evaluate the system with a variety of different bandwidth parameters τ. In particular, try $\tau = 0.01, 0.050.1, 0.51.0, 5.0$. How does the classification boundary change when varying this parameter? Can you predict what the decision boundary of ordinary (unweighted) logistic regression would look like?

~~3.~~ Multivariate least squares

So far in class, we have only considered cases where our target variable y is a scalar value. Suppose that instead of trying to predict a single output, we have a training set with multiple outputs for each example:

$$\{(x^{(i)}, y^{(i)}),\ i = 1, \ldots, m\},\ x^{(i)} \in \mathbb{R}^n,\ y^{(i)} \in \mathbb{R}^p.$$

Thus for each training example, $y^{(i)}$ is vector-valued, with p entries. We wish to use a linear model to predict the outputs, as in least squares, by specifying the parameter matrix Θ in

$$y = \Theta^T x,$$

where $\Theta \in \mathbb{R}^{n \times p}$.

(a) The cost function for this case is

$$J(\Theta) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{p} \left((\Theta^T x^{(i)})_j - y_j^{(i)} \right)^2.$$

Write $J(\Theta)$ in matrix-vector notation (i.e., without using any summations). [Hint: Start with the $m \times n$ design matrix

$$X = \begin{bmatrix} — & (x^{(1)})^T & — \\ — & (x^{(2)})^T & — \\ & \vdots & \\ — & (x^{(m)})^T & — \end{bmatrix}$$

and the $m \times p$ target matrix

$$Y = \begin{bmatrix} — & (y^{(1)})^T & — \\ — & (y^{(2)})^T & — \\ & \vdots & \\ — & (y^{(m)})^T & — \end{bmatrix}$$

and then work out how to express $J(\Theta)$ in terms of these matrices.]

(b) Find the closed form solution for Θ which minimizes $J(\Theta)$. This is the equivalent to the normal equations for the multivariate case.

(c) Suppose instead of considering the multivariate vectors $y^{(i)}$ all at once, we instead compute each variable $y_j^{(i)}$ separately for each $j = 1, \ldots, p$. In this case, we have a p individual linear models, of the form

$$y_j^{(i)} = \theta_j^T x^{(i)}, \ j = 1, \ldots, p.$$

(So here, each $\theta_j \in \mathbb{R}^n$). How do the parameters from these p independent least squares problems compare to the multivariate solution?

4. **Naive Bayes**

In this problem, we look at maximum likelihood parameter estimation using the naive Bayes assumption. Here, the input features $x_j, \ j = 1, \ldots, n$ to our model are discrete, binary-valued variables, so $x_j \in \{0, 1\}$. We call $x = [x_1 \ x_2 \ \cdots \ x_n]^T$ to be the input vector. For each training example, our output targets are a single binary-value $y \in \{0, 1\}$. Our model is then parameterized by $\phi_{j|y=0} = p(x_j = 1|y = 0)$, $\phi_{j|y=1} = p(x_j = 1|y = 1)$, and $\phi_y = p(y = 1)$. We model the joint distribution of (x, y) according to

$$\begin{aligned} p(y) &= (\phi_y)^y (1 - \phi_y)^{1-y} \\ p(x|y = 0) &= \prod_{j=1}^{n} p(x_j|y = 0) \\ &= \prod_{j=1}^{n} (\phi_{j|y=0})^{x_j} (1 - \phi_{j|y=0})^{1-x_j} \\ p(x|y = 1) &= \prod_{j=1}^{n} p(x_j|y = 1) \\ &= \prod_{j=1}^{n} (\phi_{j|y=1})^{x_j} (1 - \phi_{j|y=1})^{1-x_j} \end{aligned}$$

(a) Find the joint likelihood function $\ell(\varphi) = \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \varphi)$ in terms of the model parameters given above. Here, φ represents the entire set of parameters $\{\phi_y, \ \phi_{j|y=0}, \ \phi_{j|y=1}, \ j = 1, \ldots, n\}$.

(b) Show that the parameters which maximize the likelihood function are the same as

those given in the lecture notes; i.e., that

$$\begin{aligned}\phi_{j|y=0} &= \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}} \\ \phi_{j|y=1} &= \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}} \\ \phi_y &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}.\end{aligned}$$

(c) Consider making a prediction on some new data point x using the most likely class estimate generated by the naive Bayes algorithm. Show that the hypothesis returned by naive Bayes is a linear classifier—i.e., if $p(y = 0|x)$ and $p(y = 1|x)$ are the class probabilities returned by naive Bayes, show that there exists some $\theta \in \mathbb{R}^{n+1}$ such that

$$p(y = 1|x) \geq p(y = 0|x) \text{ if and only if } \theta^T \begin{bmatrix} 1 \\ x \end{bmatrix} \geq 0.$$

(Assume θ_0 is an intercept term.)

~~5~~. **Exponential family and the geometric distribution**

(a) Consider the geometric distribution parameterized by ϕ:

$$p(y; \phi) = (1 - \phi)^{y-1}\phi, \ y = 1, 2, 3, \ldots.$$

Show that the geometric distribution is in the exponential family, and give $b(y)$, η, $T(y)$, and $a(\eta)$.

(b) Consider performing regression using a GLM model with a geometric response variable. What is the canonical response function for the family? You may use the fact that the mean of a geometric distribution is given by $1/\phi$.

(c) For a training set $\{(x^{(i)}, y^{(i)}); \ i = 1, \ldots, m\}$, let the log-likelihood of an example be $\log p(y^{(i)}|x^{(i)}; \theta)$. By taking the derivative of the log-likelihood with respect to θ_j, derive the stochastic gradient ascent rule for learning using a GLM model with goemetric responses y and the canonical response function.