11-641: Homework 1

Due: 21 January 2016 11:59pm (Blackboard)

Name:Yan Zhao

Andrew ID: yanzhao2@andrew.cmu.edu

1: Hyperplane properties.

(a) Suppose any two points x_1, x_2 on hyperplane h, we can represent vector on h as $\vec{x_1} - \vec{x_2}$
According to hyperplane defination

$$\vec{w} \cdot (\vec{x_1} - \vec{x_2}) = w^T x_1 - w^T x_2 = b - b = 0$$

Since x_1, x_2 are any two points on h, we can prove vector $\vec{w}$ is perpendicular to hyperplane h
Then the shortest distance from the origin to hyperplane h can be defined as the absolute value of projection of vector $\vec{x}$ on vector $\vec{w}$, where vector $\vec{x}$ is vector from origin to any point on h

$$d = \frac{\vec{x} \cdot \vec{w}}{||w||} = \frac{|b|}{||w||}$$

(b) Suppose any point x_0 on hyperplane h, we can represent vector from x_0 to point x as $\vec{x} - \vec{x_0}$

Since vector $\vec{w}$ is perpendicular to hyperplane h, the projection of $x_0 \to x$ on vector $\vec{w}$ is

$$P = \frac{(\vec{x} - \vec{x_0}) \cdot \vec{w}}{||w||} = \frac{\vec{x} \cdot \vec{w} - \vec{x_0} \cdot \vec{w}}{||w||} = \frac{w^T x - b}{||w||}$$

The perpendicular distance from point x to h can be defined as the absolute value of projection of $x_0 \to x$ on vector $\vec{w}$, and projection value P will be smaller than 0 if point x is on the same side with original, so perpendicular distance d is

$$d = \frac{y(w^T x - b)}{||w||},\ y \in \{-1, 1\}$$

2: Eigenvalues and eigenvectors.

(a) For matrix A

$$A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$$

The eigenvectors $\mathbf{v}$ of transformation satisfy the equation $A\mathbf{v} = \lambda\mathbf{v}$, rearrange this equation to obtain $(A - \lambda I)\mathbf{v} = 0$. Set the determinant to zero to obtain the polynomial equation

$$p(\lambda) = |A - \lambda I| = 8 - 6\lambda + \lambda^2 = 0$$

So we can calculate the roots as $\lambda_1 = 2$ and $\lambda_2 = 4$
For $\lambda_1 = 2$, the equation becomes

$$(A - 2I)\mathbf{v_1} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Yan Zhao
yanzhao2@andrew.cmu.edu

which has the solution,

$$\mathbf{v_1} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

For $\lambda_2 = 4$, the equation becomes

$$(A - 4I)\mathbf{v_2} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

which has the solution,

$$\mathbf{v_2} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Thus, the vectors $\mathbf{v_1}$ and $\mathbf{v_2}$ are eigenvectors of A associated with the eigenvalues $\lambda_1 = 2$ and $\lambda_2 = 4$, respectively.

(b) Define the matrix P composed of eigenvectors

$$P = \begin{pmatrix} \mathbf{v_1} & \mathbf{v_2} \end{pmatrix}$$

Define diagonal matrix D composed of eigenvalues

$$D = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

Then $AP = PD$, so $A = PDP^-1$
So we can prove that

$$\begin{aligned} A^2 &= (PDP^-1)(PDP^-1) \\ &= PD(P^-1P)DP^-1 \\ &= PD^2P^-1 \end{aligned} \tag{1}$$

By induction, we can further prove $A^k = PD^kP^-1$ where

$$D^k = \begin{pmatrix} {\lambda_1}^k & 0 \\ 0 & {\lambda_2}^k \end{pmatrix}$$

So the eigenvalues of A^k are λ_1^k ,λ_2^k, the kth powers of the eigenvalues of matrix A, and that each eigenvector of A is still an eigenvector of A^k.

3: Maximum likelihood estimate

(a) For a binomial process of coin tossing

$$\begin{aligned} L(k\ headup\ out\ of\ n\ |p) &= f(k\ headup\ out\ of\ n\ |p) \\ &= \begin{pmatrix} k \\ n \end{pmatrix} \cdot p^k \cdot (1-p)^{n-k} \end{aligned} \tag{2}$$

$$\begin{aligned} F &= \ln(L(k\ headup\ out\ of\ n\ |p)) \\ &= const + k \ln p + (n-k)\ln(1-p) \end{aligned} \tag{3}$$

So for maximum likelihood,

$$\hat{p} = \arg\max_p F(p)$$

Set derivative $\frac{\partial F}{\partial p} = 0$

$$\frac{k}{p} - \frac{n-k}{1-p} = 0$$

Thus we can get $\hat{p} = \frac{k}{n}$

(b) For a multinomial process

$$L(n_1, \cdots, n_m|p) = f(n_1, \cdots, n_m|p) = \binom{n}{n_1 \cdots n_m} \cdot \prod_j {p_j}^{n_j}$$

The log-likelihood is

$$F = \ln(L(n_1, \cdots, n_m|p)) = const + \sum_j n_j \ln p_j$$

Since $n_1 + n_2 + \cdots + n_m = n$, $p_1 + p_2 + \cdots + p_m = 1$, we use Lagrange multiplier to maximize this function

$$F\prime = \sum_j n_j \ln p_j + \lambda(1 - \sum_j p_j)$$

$$\frac{\partial F\prime}{\partial p_j} = \frac{n_j}{p_j} - \lambda = 0$$

$$\frac{\partial F\prime}{\partial \lambda} = (1 - \sum_j p_j) = 0$$

So

$$n_j = \lambda \cdot p_j$$

$$\sum_j n_j = \lambda \cdot \sum_j p_j$$

$$n = \lambda$$

$$\hat{p}_j = \frac{n_j}{n}$$

References: http://www.cs.ubc.ca/ murphyk/Teaching/CS340-Fall06/reading/bernoulli.pdf

---

4: Calculus

(a) For $u = \frac{1}{1+e^{-x}}$, make $Y = e^{-x}$

$$\begin{aligned} \frac{du}{dx} &= \frac{d\frac{1}{1+Y}}{dY} \cdot \frac{dY}{dx} \\ &= -\frac{1}{(1+e^{-x})^2} \cdot (-e^{-x}) \\ &= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} \\ &= u(1-u) \end{aligned} \tag{4}$$

(b) Gradient of l

$$\begin{aligned}\nabla l &= \frac{dl}{du}\cdot\frac{du}{dz}\cdot\nabla z \\ &= (\frac{y}{u}-\frac{1-y}{1-u})\cdot u(1-u)\cdot(1,x_1,\cdots,x_m)^T \\ &= (y-u)\cdot(1,x_1,\cdots,x_m)^T \\ &= (y-u, x_1(y-u),\cdots,x_m(y-u))^T\end{aligned} \tag{5}$$

(c) Pairwise 2nd order derivative $H_{jj\prime}$

$$\begin{aligned}\nabla\nabla l &= \begin{pmatrix} \frac{\partial^2 l}{\partial w_0\partial w_0} & \frac{\partial^2 l}{\partial w_1\partial w_0} & \cdots & \frac{\partial^2 l}{\partial w_m\partial w_0} \\ \frac{\partial^2 l}{\partial w_0\partial w_1} & \frac{\partial^2 l}{\partial w_1\partial w_1} & \cdots & \frac{\partial^2 l}{\partial w_m\partial w_1} \\ \cdots & \cdots & \ddots & \cdots \\ \frac{\partial^2 l}{\partial w_0\partial w_m} & \frac{\partial^2 l}{\partial w_1\partial w_m} & \cdots & \frac{\partial^2 l}{\partial w_m\partial w_m} \end{pmatrix} \\ &= \begin{pmatrix} u(u-1) & x_1u(u-1) & \cdots & x_mu(u-1) \\ x_1u(u-1) & {x_1}^2u(u-1) & \cdots & x_1x_mu(u-1) \\ \cdots & \cdots & \ddots & \cdots \\ x_mu(u-1) & x_1x_mu(u-1) & \cdots & {x_m}^2u(u-1) \end{pmatrix}\end{aligned} \tag{6}$$

(d) The cost log-likelihood function in logistic regression is defined as l shown in previous question. By defination of concave, a differentiable function l is concave on an interval if its derivative function l' is monotonically decreasing on that interval: a concave function has a decreasing slope.

Thus for a twice-differentiable function l, if the second derivative l'' is negative, then the graph is concave.

According to previous prove, $\frac{\partial^2 l}{\partial w_j\partial w_j} = {x_j}^2u(u-1)$, which is the diagonal values of the pairwise 2nd order derivative matrix $H_{jj\prime}$, since $0 \le u \le 1$, so $\frac{\partial^2 l}{\partial w_j\partial w_j} \le 0$. Then we can prove the log-likelihood cost function of logestic regression is concave.