EE M146
 Introduction to Machine Learning
 Instructor: Lara Dolecek
 TA: Zehui (Alex) Chen, Ruiyi (John) Wu

Discussion 2
 Friday, April 12, 2019

1. Matrix calculus review

   (a) Gradient of differentiable function $f : \mathbb{R}^n \to \mathbb{R}$:
   $\nabla f(x) = \left[\frac{\partial}{\partial x_1} f(x), \frac{\partial}{\partial x_2} f(x), \cdots, \frac{\partial}{\partial x_n} f(x)\right]^T$.

   - $\nabla_w(w^T b)$
   - $\nabla_w(\|w\|^2)$
   - $\nabla_w(w^T A w)$
   - $\nabla_w(w^T X^T X w)$

   (b) Jacobian/derivative matrix of differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$:

$$\boldsymbol{J} = \begin{bmatrix} \nabla f_1(x)^T \\ \nabla f_1(x)^T \\ \vdots \\ \nabla f_m(x)^T \end{bmatrix}, \boldsymbol{J}_{ij} = \frac{\partial f_i}{\partial x_j}$$

   - Ax

- Example: transformation from polar (r, θ) to Cartesian coordinates (x, y): $x = r\cos(\theta), y = r\sin(\theta)$.

(c) Hessian matrix for twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$:
$\nabla^2 f(x)_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f(x)$.
The Hessian matrix is also the derivative matrix $\boldsymbol{J}$ of the gradient $\nabla f(x)$.

- Affine function $f(x) = a^T x + b$.

- Least squares cost: $\|Ax - b\|^2$.

- Example: $4x_1^2 + 4x_1x_2 + x_2^2 + 10x_1 + 9x_2$

2. Show that for a linearly separable data set, the maximum likelihood solution for the logistic regression model is obtained by finding a vector w whose decision boundary $w^T x = 0$ separates the classes and then taking the magnitude of w to infinity.

3. In class, we provided a probabilistic interpretation of ordinary least squares. We now try to provide a probabilistic interpretation of the weighted linear regression. Consider a model where each of the N samples is independently drawn according to a normal distribution
$$P(y_n|x_n, w) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(y_n - w^T x_n)^2}{2\sigma_n^2}\right).$$

In this model, each y_n is drawn from a normal distribution with mean $w^T x_n$ and variance σ_n^2. The σ_n^2 are **known**. Write the log likelihood of this model as a function of w. Show that finding the maximum likelihood estimate of w leads to the same answer as solving a weighted linear regression. How do σ_n^2 relate to α_n?