COMP 540 HW 1

Peiguang Wang, Xinran Zhou

Due: 1/18/2018

Part 0: Background refresher

1. Generate different distributions from uniform distribution:

   (a) Plot the histogram of a categorical distribution with probabilities [0.2,0.4,0.3,0.1].

   (b) Plot the univariate normal distribution with mean of 10 and standard deviation of 1.

   (c) Produce a scatter plot of the samples for a 2-D Gaussian with mean at [1,1] and a covariance matrix [[1,0.5],[0.5,1]]

   (d) Test your mixture sampling code by writing a function that implements an equal weighted mixture of four Gaussians in 2 dimensions, centered at $(\pm 1; \pm 1)$ and having covariance I. Estimate the probability that a sample from this distribution lies within the unit circle centered at (0.1, 0.2).

2. Prove that the sum of two independent Poisson random variables is also a Poisson random variable.

   Proof. The characteristic function of a Poisson random variable is

   $$\Phi_1(u) = e^{\lambda_1(e^{iu}-1)}$$

   Let X_1 and X_2 denote two independent Poisson random variables. Let $X = X_1 + X_2$

   Let $\Phi_1(u)$ and $\Phi_2(u)$ denote the characteristic functions of X_1 and X_2:

   $$\Phi_1(u) = e^{\lambda_1(e^{iu}-1)}$$

   $$\Phi_2(u) = e^{\lambda_2(e^{iu}-1)}$$

   Let $\Phi(u)$ denote the characteristic functions of X. Since $X = X_1 + X_2$, we have:

   $$\Phi(u) = \Phi_1(u)\Phi_2(u) = e^{\lambda_1(e^{iu}-1)}e^{\lambda_2(e^{iu}-1)}$$

   Simplify the equation above,

   $$\Phi(u) = e^{(\lambda_1+\lambda_2)(\frac{\lambda_1}{\lambda_1+\lambda_2}e^{iu}+\frac{\lambda_2}{\lambda_1+\lambda_2}e^{iu})-1}.$$

   That is

   $$\Phi(u) = e^{(\lambda_1+\lambda_2)(e^{iu}-1)}.$$

   Comparing with the characteristic function of Poisson distribution, we can see that X is also a Poisson random variable. □

3. Let X_0 and X_1 be continuous random variables. Show that if

   $$P(X_0 = x_0) = \alpha_0 e^{-\frac{(x_0-\mu_0)^2}{2\sigma_0^2}}$$

   $$P(X_1 = x_1 | X_0 = x_0) = \alpha e^{-\frac{(x_1-x_0)^2}{2\sigma^2}}$$

   there exists α_1, μ_1 and σ_1 such that

   $$P(X_1 = x_1) = \alpha_1 e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}}$$

   Write down expressions for these quantities in terms of α_0, α, μ_0, σ_0 and σ.

Solution. If X,Y are both Gaussian random variable, then

$$Y|X = x \sim N\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X), \quad \sigma_Y^2(1-\rho^2)\right)$$

where μ_X, μ_Y are mean of X and Y; σ_X^2, σ_Y^2 are variance of X and Y; ρ is the correlation coefficient between X and Y.

According to the problem, X_0, X_1 and $X_1|X_0$ are all Gaussian. So we have the following equations:

$$\begin{cases} \mu_1 + \rho\frac{\sigma_1}{\sigma_0}(x_0 - \mu_0) = x_0, for\ all\ x_0 \\ \sigma_1^2(1-\rho^2) = \sigma^2 \end{cases}$$

Solve the equation, then $\sigma_1^2 = \sigma^2 + \sigma_0^2$, $\mu_1 = -\mu_0$. And since $\alpha_1 = \frac{1}{\sqrt{2\pi}\sigma_1}$, we have

$$\alpha_1 = \sqrt{\frac{1}{(1/\alpha)^2 + (1/\alpha_0)^2}}$$

.

4. Find the eigenvalues and eigenvectors of the following 2×2 matrix A.

$$A = \begin{pmatrix} 13 & 5 \\ 2 & 4 \end{pmatrix}$$

Solution. Let λ and $\boldsymbol{x}$ denote the eigenvalue and eigenvector of A. According to the definition of eigenvalue,

$$A\boldsymbol{x} = \lambda\boldsymbol{x}$$

Solve the equation to get eigenvalues

$$|A - \lambda I| = 0$$

That is,

$$\lambda^2 - 14\lambda + 42 = 0$$

A has two eigenvalues: $\lambda_1 = 14$, $\lambda_2 = 3$.

When $\lambda = 14$,

$$(A - \lambda I)\boldsymbol{x} = \begin{pmatrix} -1 & 5 \\ 2 & -10 \end{pmatrix}\boldsymbol{x} = 0$$

$$\boldsymbol{x} = \begin{pmatrix} 5 & 1 \end{pmatrix}^T$$

When $\lambda = 3$,

$$(A - \lambda I)\boldsymbol{x} = \begin{pmatrix} 10 & 5 \\ 2 & 1 \end{pmatrix}\boldsymbol{x} = 0$$

$$\boldsymbol{x} = \begin{pmatrix} 1 & -2 \end{pmatrix}^T$$

In summary, A has two eigenvalues, $\lambda_1 = 14$, $\lambda_2 = 3$. The corresponding eigenvectors are $\boldsymbol{x_1} = \begin{pmatrix} 5 & 1 \end{pmatrix}^T$ and $\boldsymbol{x_2} = \begin{pmatrix} 1 & -2 \end{pmatrix}^T$.

5. Provide one example for each of the following cases, where A and B are 2 2 matrices.

   (a) $(A+B)^2 \neq A^2 + 2AB + B^2$

   (b) $AB = 0, A \neq 0, B \neq 0$

Solution. (a) one example that satisfies (a) is:

$$A = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, B = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

Calculate left,

$$left = (A+B)^2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$$

Calculate right,

$$right = A^2 + 2AB + B^2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} + \mathbf{0} + \mathbf{0} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

And $left \neq right$

(b) one example that satisfies (b) is:

$$A = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, B = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

where $A \neq 0$, and $B \neq 0$. Calculate AB,

$$AB = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \mathbf{0}$$

6. Let u denote a real vector normalized to unit length. That is, $u^T u = 1$. Show that

$$A = I - 2uu^T$$

is orthogonal, i.e., $A^T A = 1$.

Proof. Derive from left,

$$A^T A = (I - 2uu^T)^T (I - 2uu^T) = (I - 2uu^T)(I - 2uu^T) = I - 2uu^T - 2uu^T + 4uu^T = I$$

So $left = right$. □

Part 1: Locally weighted linear regression

1. Show that $J(\theta)$ can be written in the form

$$J(\theta) = (X\theta - y)^T W (X\theta - y)$$

for an appropriate diagonal matrix W, where X is the $m \times d$ input matrix and y is a $m \times 1$ vector denoting the associated outputs. State clearly what W is.

Proof. We know that $J(\theta)$ can also be written as

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} w^{(i)} (\theta^T x^{(i)} - y^{(i)})^2$$

where $x^{(i)}$ is $d \times 1$ vector and θ is a $d \times 1$ vector. We consider each row of the matrix X as as a $1 \times d$vector x^i, so we can write $X = [x^1, x^2, ..., x^m]^T$. So

$$J(\theta) = (X\theta - y)^T W (X\theta - y) = [x^1\theta - y^1, x^2\theta - y^2, ..., x^m\theta - y^m] W [x^1\theta - y^1, x^2\theta - y^2, ..., x^m\theta - y^m]^T$$

So W is a $m \times m$ diagonal matrix

$$W = \begin{pmatrix} 2w^{(1)} & 0 & 0 & 0 \\ 0 & 2w^{(2)} & 0 & 0 \\ 0 & 0 & ... & 0 \\ 0 & 0 & 0 & 2w^{(m)} \end{pmatrix}$$

□

2. If all the $w^{(i)}$'s are equal to 1, the normal equation to solve for the parameter θ is:

$$X^T X\theta = X^T y$$

and the value of θ that minimizes $J(\theta)$ is $(X^T X)^{-1}X^T y$. By computing the derivative of the weighted $J(\theta)$ and setting it equal to zero, generalized the normal equation to the weighted setting and solve for θ in closed form in terms of W, X and y.

Proof.

$$J(\theta) = (X\theta - y)^T W(X\theta - y) = \theta^T X^T W X\theta^T - \theta^T X^T W y - y^T W X\theta + y^T W y$$

Compute the derivative of $J(\theta)$

$$\frac{\partial J(\theta)}{\partial(\theta)} = 2X^T W X\theta - X^T W y - X^T W^T y$$

Since W is a diagonal matirx $W = W^T$, the equation can be written as

$$\frac{\partial J(\theta)}{\partial(\theta)} = 2X^T W X\theta - 2X^T W y$$

By setting it equal to zero, we can find the value of θ that minimizes $J(\theta)$, the equation is :

$$X^T W X\theta = X^T W y$$

So the value of θ in form in terms of W, X and y is $(X^T W X)^{-1}X^T W y$. □

3. To predict the target value for an input vector x, one choice for the weighting function $w^{(i)}$ is:

$$w^{(i)} = \exp(-\frac{(x - x^{(i)})^T(x - x^{(i)})}{2\tau^2})$$

Points near x are weighted more heavily than points far away from x. The parameter τ is a band width defining the sphere of influence around x. Note how the weights are defined by the input x. Write down an algorithm for calculating θ by batch gradient descent for locally weighted linear regression. Is locally weighted linear regression a parametric or a non-parametric method?

Part 2: Properties of the linear regression estimator

1. Show that $E[\theta] = \theta^*$ for the least squares estimator.

Proof. In part 1 problem 2, we can get thee value of θ given the normal equation $X^T X\theta = X^T y$ is

$$\theta = (X^T X)^{-1}X^T y$$

The date comes from the linear model:

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

the expectation of θ is

$$\begin{aligned}
E[\theta] &= E[(X^T X)^{-1}X^T y] \\
&= E[(X^T X)^{-1}X^T(X\theta^* + \epsilon)] \\
&= E[(X^T X)^{-1}(X^T X\theta^* + X^T\epsilon)] \\
&= E[(X^T X)^{-1}X^T X\theta^* + (X^T X)^{-1}X^T\epsilon] \\
&= E[\theta^*] + E[(X^T X)^{-1}X^T\epsilon]
\end{aligned}$$

since each $\epsilon^{(i)}$is an independent random variable drawn from a normal distribution with zero mean and variance σ^2 and θ^* is a true parameter that has certain value. Then $E[\theta]$ can be written as

$$E[\theta] = E[\theta^*] + 0 = \theta^*$$

□

2. Show that the variance of the least squares estimator is $Var(\theta) = (X^T X)^{-1}\sigma^2$.

Proof.

$$Var(\theta) = E[\theta^2] - (E[\theta])^2$$

since we already knew that $E[\theta] = \theta^*$. So in order to get $Var(\theta)$, all we need to do is to compute $E[\theta^2]$.

$$\begin{aligned} E[\theta^2] &= E[((X^T X)^{-1} X^T y)((X^T X)^{-1} X^T y)^T] \\ &= E[(\theta^* + (X^T X)^{-1} X^T \Sigma)(\theta^* + (X^T X)^{-1} X^T \Sigma)^T] \\ &= E[\theta^* \theta^{*T} + (X^T X)^{-1} X^T \Sigma \theta^{*T} + \theta^* \Sigma^T X (X^T X)^{-1} + (X^T X)^{-1} X^T \Sigma \Sigma^T X (X^T X)^{-1}] \end{aligned}$$

Σ is the covariance matrix generated by ϵ and each $\epsilon^{(i)}$is an independent random variable drawn from a normal distribution with zero mean and variance σ^2. Therefore the expectation of Σ is zero. Σ is also independent to X and θ^*, $\Sigma = \sigma^2 I$, where I is the identity matrix. Therefore

$$E[\theta^2] = (\theta^*)^2 + \sigma^2 I (X^T X)^{-1}$$

Then we have

$$\begin{aligned} Var(\theta) &= E[\theta^2] - (E[\theta])^2 \\ &= (\theta^*)^2 + \sigma^2 I (X^T X)^{-1} - (\theta^*)^2 \\ &= (X^T X)^{-1} \sigma^2 \end{aligned}$$

□