Statistical Machine Learning
Assignment One: Model Selection, Probability Theory and Distributions

Author : Sagar Kukreja

1) Show that the variance of a sum is var[X+Y] = var[X]+var[Y]+2cov[X,Y].

Var[X] = E[X^2] - $E[X]^2$

Similarly we can write,

Var[X+Y] = E[$(X+Y)^2$] - $(E[X+Y])^2$

= E[$X^2 + Y^2 + 2XY$] - $(E[X] + E[Y])^2$

= E[X^2] + E[Y^2] + 2E[XY] - E[X]^2 - E[Y]^2 - 2E[X]E[Y]

= E[X^2] - E[X]^2 + E[Y^2] - E[Y]^2 + 2E[XY] - 2E[X]E[Y]

= var[X] + var[Y] + 2 COV[X,Y] (COV[X,Y] = E[XY] - E[X]E[Y])

2) Suppose Θ ~ Beta(a, b) , derive the mean, mode and variance.

Beta(Θ | a, b) = $\frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\Theta^{a-1}(1-\Theta)^{b-1}$

Mean --

E(Θ) = $\int_0^1 \Theta$ .Beta(Θ | a, b)

$= \int_0^1 \Theta \, . \, \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\Theta^{a-1}(1-\Theta)^{b-1}$

$= \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\Theta^{a+1-1}(1-\Theta)^{b-1}$

(as $\int_0^1 Beta(\Theta | a, b) = 1 \Rightarrow \int_0^1 \Theta^{a+1-1}(1-\Theta)^{b-1} = \frac{\Gamma(a+1)\,\Gamma(b)}{\Gamma(a+1+b)}$ when compared to Beta distribution)

$= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} * \frac{\Gamma(a+1)\,\Gamma(b)}{\Gamma(a+1+b)}$

$= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} * \frac{a\,.\,\Gamma(a)\,\Gamma(b)}{(a+b)\,.\,\Gamma(a+b)}$ (as $\Gamma(x+1) = x\Gamma x$)

E[Θ] = $\frac{a}{a+b}$

Variance --

var[Θ] = E[Θ^2] - $(E[\Theta])^2$

E[Θ^2] = $\int_0^1 \Theta^2$ .Beta(Θ | a, b)

$$= \int_0^1 \Theta^2 \cdot \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} \Theta^{\,a-1} (1-\Theta)^{b-1}$$

$$= \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} \Theta^{\,a+2-1}\ (1-\Theta)^{b-1}$$

(as $\int_0^1 Beta(\Theta|\, a, b) = 1 \Rightarrow \int_0^1 \Theta^{\,a+1-1}\ (1-\Theta)^{b-1} = \frac{\Gamma(a+2)\,\Gamma(b)}{\Gamma(a+2+b)}$ when compared to Beta distribution)

$$= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} * \frac{\Gamma(a+2)\,\Gamma(b)}{\Gamma(a+2+b)}$$
$$= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} * \frac{a\ .\ (a+1).\Gamma(a)\,\Gamma(b)}{(a+b+1).(a+b)\ .\ \Gamma(a+b)} \quad \text{(as } \Gamma(x+1) = x\Gamma x \text{)}$$
$$= \frac{a.(a+1)}{(a+b+1).(a+b)}$$

$(E[\Theta])^2 = \frac{a^2}{(a+b)^2}$

var[Θ] = E[Θ^2] - $(E[\Theta])^2$

$$= \frac{a.(a+1)}{(a+b+1).(a+b)} - \frac{a^2}{(a+b)^2} = \frac{ab}{(a+b)^2(a+b+1)}$$

Mode -- The mode is where pdf reaches its maximum, hence we differentiate the pdf and set it to zero.

Beta(Θ | a, b) = $\frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} \Theta^{\,a-1} (1-\Theta)^{b-1}$

Differentiating w.r.t Θ and setting it to zero :

$$\frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} ((a-1) * \Theta^{\,(a-2)} * (1-\Theta)^{b-1} + (b-1) * (-1) * (1-\Theta)^{(b-2)} * \Theta^{\,a-1}) = 0$$

On solving it gives ::

$\Rightarrow (a-1) * (1-\Theta) - (\Theta * (b-1)) = 0$

$\Rightarrow a - a\Theta - 1 + \Theta - \Theta b + \Theta = 0$

$\Rightarrow Mode = \frac{(a-1)}{(a+b-2)}$

3) Since a positive definite matrix Σ can be defined as the quadratic form $U^T \Lambda U$, show that a necessary and sufficient condition for Σ to be positive definite is that all the eigenvalues λ_i of Λ are positive.

For Σ to be positive definite , all the eigenvalues λ_i of Λ should be positive:

1) Trace(sum of diagonal elements) of Λ will be positive and $|\Lambda|$ must be positive to be positive definite as Λ is positive definite.
2) product of eigen-values = $|\Lambda|$ which will be positive

From above 2 statements we can conclude that necessary and sufficient condition for Σ to be positive definite is that λ_i should be positive.

4) Derive the maximum likelihood solutions for the mean and the variance of a univariate Gaussian distribution by maximize the log likelihood function with respect to Σ and μ.

Let $X_1, X_2, ...X_n$ be i.i.d random variables and let X_i be the value each X_i takes. The density for each X_i is

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\Pi}\ .\ \sigma} e^{\frac{-(x_i-\mu)^2}{2.\sigma^2}}$$

Since , the X_i are independent , their joint pdf is the product of individual pdf's:

$$f(x_1,\ x_2, ...x_n\ |\mu,\ \sigma) = \left(\frac{1}{\sqrt{2\Pi}\ .\ \sigma}\right)^n e^{-\Sigma^n{}_{i=1} \frac{(x_i-\mu)^2}{2.\sigma^2}}$$

For the fixed data $x_1,\ x_2, ...x_n$, the likelihood and log likelihood are :

$$f(x_1,\ x_2, ...x_n\ |\mu,\ \sigma) = \left(\frac{1}{\sqrt{2\Pi}\ .\ \sigma}\right)^n e^{-\Sigma^n{}_{i=1} \frac{(x_i-\mu)^2}{2.\sigma^2}}$$

And

$$ln(f(x_1,\ x_2, ...x_n\ |\mu,\ \sigma)) = -n\ ln\ (\sqrt{2\Pi}\ ) - n\ ln\ (\sigma) - \Sigma^n{}_{i=1} \frac{(x_i-\mu)^2}{2.\sigma^2}$$

Since $ln(f(x_1,\ x_2, ...x_n\ |\mu,\ \sigma))$ is a function of 2 variables, we use partial derivative to find the maximum likelihood

$$\frac{df(x_1, x_2, ...x_n\ |\mu, \sigma)}{d\mu} = \Sigma^n{}_{i=1} \frac{(x_i-\mu)}{\sigma^2} = 0$$
$$\Rightarrow \Sigma^n{}_{i=1} x_i = n\mu$$

$$\Rightarrow \mu = \frac{\Sigma^{n}_{i=1} x_i}{n} = mean(x)$$

To find σ we differentiate and solve with respect to σ :

$$\frac{df(x_1, x_2, ... x_n \mid \mu, \sigma)}{d\sigma} = -\frac{n}{\sigma} + \Sigma^{n}_{i=1} \frac{(x_i - \mu)^2}{\sigma^3} = 0$$

$$\Rightarrow \sigma^2 = \Sigma^{n}_{i=1} \frac{(x_i - \mu)^2}{n}$$

$$\Rightarrow \sigma^2 = \frac{1}{n} \Sigma^{n}_{i=1} (x_i - \mu)^2 \text{ = variance of data}$$

5) Write a pseudo-code for using cross-validation to determine the best K value of K nearest neighbors classifier.

1) Repeat the following M (for M-fold cross-validation) times
   - Randomly split the data into two sets (train and test). Put the $\frac{M-1}{M}$ percent of the data into train and the remaining $\frac{1}{M}$ percent of the data into test.
   - For each value of k we are interested in
     - Fit the model on train.
     - Compute the error on the test set

2) Now we have a k x M matrix of errors. For each value of the tuning parameter k compute the average cv-error across the M folds.

3) Select the value of k with the best cross validation error.