CISC 863 Statistical Machine Learning

Assignment One: Probability Theory and Distributions

Show that the variance of a sum is $\mathrm{var}[X+Y] = \mathrm{var}[X] + \mathrm{var}[Y] + 2\,\mathrm{cov}[X,Y]$.

Ans: $\mathrm{var}[X+Y] = E[(X+Y-\mu_x-\mu_y)^2]$

$\because$ we know that, $\mathrm{var}[x] = E(x-\mu)^2$

$= E[((X-\mu_x)+(Y-\mu_y))^2]$

$= E[(X-\mu_x)^2 + (Y-\mu_y)^2 + 2(X-\mu_x)(Y-\mu_y)]$

$= E[(X-\mu_x)^2] + E[(Y-\mu_y)]^2 + E[2(X-\mu_x)(Y-\mu_y)]$

$= \mathrm{var}[X] + \mathrm{var}[Y] + 2\,\mathrm{cov}[X,Y]$ (showed) most of the pr...

Suppose $\theta \sim Beta(\alpha, \beta)$, derive the mean, mode and variance.

Mean: From given eqⁿ of expected value of mean we can write that,

$$E[X] = \int_0^1 x \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} dx \quad \ldots\ldots 1$$

$$= \frac{1}{B(\alpha\,\beta)} \left[\frac{x^{\alpha}(-1)(1-x)^{\beta}}{\beta} \Big|_0^1 - \int_0^1 \alpha x^{\alpha-1} (-1) \frac{(1-x)^{\beta}}{\beta} dx \right]$$

$$= \frac{1}{B(\alpha,\beta)} \int_0^1 \frac{\alpha}{\beta} x^{\alpha-1} (1-x)^{\beta-1} (1-x)\, dx$$

$$= \frac{1}{B(\alpha,\beta)} \left[\int_0^1 \frac{\alpha}{\beta} x^{\alpha-1} (1-x)^{\beta-1} dx - \int_0^1 x \frac{\alpha}{\beta} x^{\alpha-1} (1-x)^{\beta-1} dx \right]$$

$$= \frac{\alpha}{\beta} \left(1 - E[x]\right)$$

$$\therefore E[X] = \frac{\alpha}{\alpha+\beta} = \text{Expected value of mean.}$$

Mode: In order to calculate mode, we have to differentiate the Beta distribution and equal it to zero.

$$\text{mode} = \frac{d}{dx} Beta(\alpha,\beta) = 0 \Rightarrow \frac{d}{dx}\left(x^{\alpha-1}(1-x)^{\beta-1}\right) = 0$$

$$\cancel{\Rightarrow (\alpha-1) x^{\alpha-2} (\beta-1)(1-x)^{\beta-2}} \Rightarrow (\alpha-1) x^{\alpha-2} (1-x)^{\beta-1} - x^{\alpha-1} (\beta-1)(1-x)^{\beta-2} = 0$$

$$\Rightarrow (\alpha-1)(1-x) - x(\beta-1) = 0 \Rightarrow (\alpha-1) - x(\alpha-1+\beta-1) = 0$$

$$\Rightarrow x = \frac{\alpha-1}{\alpha+\beta-2}.$$

__Variance:__ We know that,

$$\mathrm{var}[X] = E[X^2] - E[X]^2$$

We know, $E(X) = \frac{\alpha}{\alpha+\beta}$.

Thus we need $E[X^2]$ to calculate.

$$E[X^2] = \int_0^1 x^2 \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\,dx$$

$$= \frac{1}{\beta\,B(\alpha,\beta)}\left[-x^{\alpha+1}(1-x)^{\beta}\Big|_0^1 - \int_0^1 -(\alpha+1)\,x^{\alpha}(1-x)^{\beta}\,dx\right]$$

$$= \frac{\alpha+1}{\beta}\int_0^1 \frac{1}{B(\alpha,\beta)}\, x\, x^{\alpha-1}.(1-x)^{\beta}.$$

$$= \frac{\alpha+1}{\beta}\left[\int_0^1 \frac{1}{B(\alpha,\beta)}\, x\, x^{\alpha-1}.(1-x)^{\beta-1} - \int_0^1 \frac{1}{B(\alpha,\beta)}\, x^2 x^{\alpha-1}(1-x)^{\beta-1}\right]$$

$$= \frac{\alpha+1}{\beta}\left[E[X] - E[X^2]\right] = \frac{\alpha+1}{\beta}\left[\frac{\alpha}{\alpha+\beta} - E[X^2]\right]$$

$$\therefore\ E(X^2) = \frac{\alpha+1}{\alpha+\beta+1}\cdot\frac{\alpha}{\alpha+\beta}.$$

$$\therefore\ \mathrm{var}(X) = \frac{\alpha+1}{\alpha+\beta+1}\cdot\frac{\alpha}{\alpha+\beta} - \frac{\alpha^2}{(\alpha+\beta)^2}.$$

$$= \frac{\alpha^2+\alpha}{(\alpha+\beta)(\alpha+\beta+1)} - \frac{\alpha^2}{(\alpha+\beta)^2}.$$

$$= \frac{(\alpha^2+\alpha)(\alpha+\beta) - \alpha^2(\alpha+\beta+1)}{(\alpha+\beta+1)(\alpha+\beta)^2}$$

$$= \frac{\cancel{\alpha^3} + \cancel{\alpha^2} + \alpha^2\beta + \alpha\beta - \cancel{\alpha^3} - \alpha^2\beta - \cancel{\alpha^2}}{(\alpha+\beta+1)(\alpha+\beta)^2}$$

$$= \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}.$$

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.

~~Sufficient~~

As we know that,

$\Sigma = U^T \Lambda U$ will be positive definite if $U' \Lambda U > 0$

Now, $$U^T \Lambda U = [x_1 \cdots x_N] \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$

$$= x_1^2 \lambda_1 + x_2^2 \lambda_2 + \cdots + x_N^2 \lambda_N \quad \cdots (1)$$

(1) eqn needs to be greater than zero to make Σ positive definite.

As we know x_i^2 must be greater than zero.

Thus we need to make sure that all eigen values λ_i ~~must~~ ~~be~~ need to be ~~zero~~ positive to make Σ positive definite.

Proof by contradiction:

Say one eigenvalue λ_i is zero then there will exist at least one eigenvector e such that $\Sigma e = 0$. So $e^T \Sigma e = 0$ and condition ~~$x^T M x > 0$~~. $U^T \Lambda U > 0$ doesn't hold when $x = e$.

Another condition we may assume that if the matrix Σ negative i.e. $\lambda < 0$. Then there will exist at least one eigenvector e such that $\Sigma e = \lambda e$. So $e^T \Sigma e = e^T \lambda e = \lambda e^T e = \lambda |e|^2$.

Since $\lambda < 0$, condition $x^T M x > 0$ does not hold when $x = e$ so, it is necessary for the eigenvalues of a positive definite matrix to be positive.

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 μ and Σ.

Maximum likelihood solution for mean;

We know that, likelihood is,

$$P(x_1, x_2 \cdots x_N | \mu) = \prod_{i=1}^{N} P(x_i | \mu)$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2}$$ [∵ this is a gaussian distribution.

Now since this is a guassian Univariate distribution, we can maximize the log likelihood.

$$\cancel{P(x|\mu)} \quad \log(P(x_1, x_2 \cdots x_N | \mu)) = \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x_i-\mu)^2}{2\sigma^2} \quad \text{--- (1)}$$

we take the derivative with respect to, μ,

$$\frac{d \log(P(x_1, x_2 \cdots x_N | \mu))}{d\mu} = \sum_{i=1}^{N} \frac{(x_i - \mu)}{\sigma^2}.$$

Since we want to maximize assume that left side is '0';

$$0 = \sum_{i=1}^{N} \frac{(x_i - \mu)}{\sigma^2}$$

$$\Rightarrow 0 = \sum_{i=1}^{N} (x_i - \mu)$$

$$\Rightarrow \sum_{i=1}^{N} \mu = \sum_{i=1}^{N} x_i$$

$$\Rightarrow N\mu = \sum_{i=1}^{N} x_i$$

$$\therefore \hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{N}. \quad \text{Ans.}$$

In the same way we can take derivative of ① with respect to σ^2,

$$\frac{d \log\left(p(x_1, x_2, \dots x_N \mid \mu)\right)}{d\sigma^2} = \frac{d}{d\sigma^2}\left[\sum_{i=1}^{N}\left\{\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right\}\right]$$

$$= \frac{d}{d\sigma^2}\left\{-\frac{N}{2}\log\left(2\pi\sigma^2\right)\right\} - \frac{1}{2}\sum (x_i-\mu)^2 \, \frac{d}{d\sigma^2}\frac{1}{\sigma^2}.$$

$$= \cancel{\frac{1}{2}\log 2\pi} - \frac{N}{2}\frac{d}{d\sigma^2}\log\sigma^2 + \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^2(\sigma^2)^{-2}$$

$$= \cancel{-\frac{1}{2}\log(2\pi)} - \frac{1}{2}\frac{N}{\sigma^2} + \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^2\frac{1}{(\sigma^2)^2}$$

To maximize we can make left hand side zero,

$$0 = -\frac{1}{2}\frac{N}{\sigma^2} + \frac{1}{2}\sum_{i=1}^{N}(x_1-\mu)\frac{1}{(\sigma^2)^2}$$

$$\Rightarrow \frac{N}{\sigma^2} = \cancel{\frac{1}{2}}\sum_{i=1}^{N}(x_i-\mu)^2\frac{1}{(\sigma^2)^2}$$

$$\Rightarrow \frac{\sigma^2}{N} = \frac{(\sigma^2)^2}{\sum_{i=1}^{N}(x_i-\mu)^2}$$

$$\Rightarrow \frac{1}{N} = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i-\mu)^2} \quad [\because \sigma^2 \neq 0]$$

$$\Rightarrow \sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} \,. \text{ Ans.}$$

Plot Gaussian likelihoods with unknown means, conjugate priors of the means, and their corresponding posterior distributions with tools you are comfortable with (e.g., Matlab, R, Octave) and different parameter settings

- *Prior*: $N(\mu|0,6)$
- *Likelihood*: $N(D|\mu, 10)$ (data file in mycourses)
- *Posterior*: $N(\mu|D, 0,6)$
- *Predictive*: $N(x^* = 2.4|D)$

For posterior we know,

$$\sigma_N^2 = \frac{1}{\left(\frac{N}{\sigma^2}\right) + \left(\frac{1}{\sigma_0^2}\right)} = \frac{1}{\left(\frac{20}{10}\right) + \left(\frac{1}{6}\right)} = \left(\frac{1}{2 + 1/6}\right) = \frac{6}{13}$$

$$\mu_N = \left(\frac{N\bar{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) \cdot \sigma_N^2 = \left(\frac{20}{\sigma^2} \cdot \bar{x} + 0\right) \cdot \frac{6}{13} = \left(\frac{20}{10}\bar{x}\right) \cdot \frac{6}{13}.$$

For predictive posterior,

$$\sigma_{NP}^2 = \sigma_N^2 + \sigma^2 = \frac{6}{13} + 10 = \frac{136}{13}.$$

μ_N = same as posterior.

O Scalar QDA: Consider the following training set of heights x (in inches) and gender y (male/female) of some college students: $x=\{67,79,71,68,67,60\}$, and $y=\{m,m,m,f,f,f\}$.

a. Fit a Bayes classifier to this data, using Maximum Likelihood estimation (MLE), i.e., estimate the parameters of the class conditional likelihoods

$$p(x \mid y=c)=N(x \mid \mu_c,\sigma_c)$$

And the class prior

$$p(y=c)=\pi_c$$

What are your values of μ_c, σ_c, π_c for $c='m','f'$? Show your work (so you can get partial credit if you make an arithmetic error).

b. Compute $p(y='m' \mid x,\hat{\theta})$, where $x=72$, and $\hat{\theta}$ are the MLE parameters. (This is called a plug-in prediction.)

Note: solve this exercise by hand AND using a computer (Matlab, Python, Octave, whatever). Show your work (derivation and code).

Hint: refer to the Probability Theory slides on QDA: 99-104

Ans: Here given, heights, $x = \{67,79,71,68,67,60\}$
$y = \{m,m,m,f,f,f\}$

$p(x=m) = 0.5 = \pi_m$ $\quad\quad \mu_m = 72.3$

$p(x=f) = 0.5 = \pi_f$ $\quad\quad \mu_f = 65.0$

As we know, variance,

$$\sigma_m = \frac{\Sigma(x_m-\mu_m)^2}{\Sigma m} = 4.98$$

$$\sigma_f = \frac{\Sigma(x_f-\mu_f)^2}{\Sigma f} = 3.56$$

$$p(y='m' \mid \underset{72}{x},\hat{\theta}) = \frac{p(x=72 \mid y=m) \times prior(x=m)}{P_r(x=72 \mid y=m) \times prior(x=m) + P_r(x=72 \mid y=f) \times prior(y=f)}$$

$= 0.83.$ Ans.

Scanned by CamScanner