

2016 Spring Notes

Yue Yu

February 20, 2016

Contents

1	各种相关背景知识	2
1.1	概率论	2
1.1.1	条件概率期望:	2
1.1.2	Correlation coefficient(相关系数)	2
1.1.3	协方差矩阵	2
1.1.4	Multivariate normal distribution	2
1.1.5	Laplace Distribution:	3
1.2	矩阵求导法则 :	3
2	Information Theory	4
2.1	Differential Entropy	4
2.1.1	常见 Differential Entropy	4
2.1.2	Properties	4
3	Machine Learning	5
3.1	MAXIMUM LIKELIHOOD ESTIMATION (MLE)	5
3.1.1	GAUSSIAN MLE	5
3.2	Linear Regression	5
3.2.1	Least Squares	5
3.3	Classification	5

1 各种相关背景知识

1.1 概率论

1.1.1 条件概率期望:

$$E(X) = E(E(X|Y)) \quad (1)$$

1.1.2 Correlation coefficient(相关系数)

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad (2)$$

$$= \frac{E(X - E(X))(Y - E(Y))}{\sigma_X \sigma_Y} \quad (3)$$

$$= \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y} \quad (4)$$

1.1.3 协方差矩阵

$$X = [X_1, \dots, X_n]^T \quad (5)$$

$$\Sigma = E[(X - E(X))(X - E(X))^T] \quad (6)$$

$$\Sigma_{i,j} = Cov(X_i, X_j) \quad (7)$$

$$= E[(X_i - \mu_i)(X_j - \mu_j)] \quad (8)$$

1.1.4 Multivariate normal distribution

- 多维高斯联合分布:

$$f_X(x_1, \dots, x_n) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$$

- 当为二维高斯分布时 (ρ 为相关系数):

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)^2} \left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} \right]\right)$$

条件概率服从:

$$P(X_1|X_2 = x_2) \sim N(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho(x_2 - \mu_2), (1 - \rho^2)\sigma_1^2)$$

1.1.5 Laplace Distribution:

- 概率密度函数

$$f(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$$

- 期望

$$\mu$$

- 方差

$$2b^2$$

1.2 矩阵求导法则：

$$\frac{\partial \mathbf{u}^T \mathbf{A} \mathbf{v}}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}} \mathbf{A} \mathbf{v} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}} \mathbf{A}^T \mathbf{u} \quad (9)$$

$$\frac{\partial (\mathbf{u}(\mathbf{x}) + \mathbf{v}(\mathbf{x}))}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}(\mathbf{x})}{\partial \mathbf{x}} + \frac{\partial \mathbf{v}(\mathbf{x})}{\partial \mathbf{x}} \quad (10)$$

$$\frac{\partial \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}^T \quad (11)$$

$$\frac{\partial \mathbf{a}}{\partial \mathbf{x}} = \mathbf{0} \quad (12)$$

$$\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{b}}{\partial \mathbf{x}} = \mathbf{A} \mathbf{b} \quad (13)$$

$$\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T) \mathbf{x} \quad (14)$$

$$= 2\mathbf{A} \mathbf{x} \quad \text{如果 } \mathbf{A} \text{ 为对称阵} \quad (15)$$

$$\frac{\partial |\mathbf{X}|}{\partial \mathbf{X}} = |\mathbf{X}| (\mathbf{X}^{-1})^T \quad (16)$$

$$\frac{\partial \ln |\mathbf{X}|}{\partial \mathbf{X}} = (\mathbf{X}^{-1})^T \quad (17)$$

2 Information Theory

2.1 Differential Entropy

2.1.1 常见 Differential Entropy

- *Uniform distribution (from 0 to a)*

$$h(X) = \log(a)$$

- *Normal Distribution*

$$X \sim N(0, \sigma^2)$$

$$h(X) = \frac{1}{2} \log 2\pi e \sigma^2$$

- *Multivariate Normal Distribution*

$$N_n \sim (\mu, K)$$

μ is mean and K is covariance matrix

$$h(X_1, \dots, X_n) = \frac{1}{2} \log(2\pi e)^n |K|$$

2.1.2 Properties

- $h(X + c) = h(X)$

- $h(aX) = h(X) + \log |a|$

- $h(AX) = h(X) + \log |\det(A)|$

3 Machine Learning

3.1 MAXIMUM LIKELIHOOD ESTIMATION (MLE)

3.1.1 GAUSSIAN MLE

$$\hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^n x_i$$
$$\hat{\Sigma}_{ML} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu}_{ML})(x_i - \hat{\mu}_{ML})^T$$

3.2 Linear Regression

3.2.1 Least Squares

Usually, for linear regression (and classification) we include an intercept term w_0 that doesn't interact with any element in the vector x . It will be convenient to attach a 1 to the first dimension of each vector x_i .

$$x_i = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{id} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1d} \\ 1 & x_{21} & \dots & x_{2d} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nd} \end{bmatrix}$$

这种情况下，得到的解为：

$$w_{ML} = (X^T X)^{-1} X^T y$$

预测新的点为：

$$y_{\text{new}} = x_{\text{new}}^T w_{ML}$$

3.3 Classification