2016 Spring Notes

Yue Yu

February 20, 2016

Contents

1 各种相关背景知识

1.1 概率论

1.1.1 条件概率期望:

$$E(X) = E(E(X|Y)) \tag{1}$$

1.1.2 Correlation coefficient(相关系数)

$$\begin{aligned} \rho_{X,Y} &= \frac{Cov(X,Y)}{\sigma_X \sigma_Y} && (2)\\ &= \frac{E(X-E(X))(Y-E(Y))}{\sigma_X \sigma_Y} && (3)\\ &= \frac{E(XY)-E(X)E(Y)}{\sigma_X \sigma_Y} && (4) \end{aligned}$$

1.1.3 协方差矩阵

$$\begin{aligned} \mathrm{X} &= [X_1, \ldots, X_n]^T && (5)\\ \Sigma &= \mathrm{E}\Big[(\mathrm{X}-\mathrm{E}(\mathrm{X}))(\mathrm{X}-\mathrm{E}(\mathrm{X}))^T\Big] && (6)\\ \Sigma_{i,j} &= Cov(X_i, X_j) && (7)\\ &= \mathrm{E}\Big[(X_i-\mu_i)(X_j-\mu_j)\Big] && (8) \end{aligned}$$

1.1.4 Multivariate normal distribution

- 多维高斯联合分布:

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

- 当为二维高斯分布时 (ρ 为相关系数) :

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

  条件概率服从:

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

1.1.5 Laplace Distribution:

- 概率密度函数

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

- 期望

$$\mu$$

- 方差

$$2b^2$$

1.2 矩阵求导法则：

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

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

$$\frac{\partial \mathbf{A x}}{\partial \mathbf{x}} = \mathbf{A^T} \tag{11}$$

$$\frac{\partial \mathbf{a}}{\partial \mathbf{x}} = \mathbf{0} \tag{12}$$

$$\frac{\partial \mathbf{x^T A b}}{\partial \mathbf{x}} = \mathbf{A b} \tag{13}$$

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

$$= \mathbf{2Ax} \quad \text{如果 A 为对称阵} \tag{15}$$

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

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

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

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

2.1.2 Properties

- $h(X + c) = h(X)$
- $h(aX) = h(X) + \log |a|$
- $h(AX) = h(X) + \log \left|\det(A)\right|$

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.

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

这种情况下，得到的解为:

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

预测新的点为:

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

3.3 Classification

3.3.1 Bayes Classifier

The Bayes classifier has the smallest prediction error of all classifiers.
假设 (X, Y) 独立同分布，那么 optimal classifier is:

$$f(x) = \arg\max_{y \in \mathcal{Y}} P(Y = y | X = x) \tag{18}$$

Using Bayes rule and ignoring $P(X = x)$ we equivalently have:

$$f(x) = \arg\max_{y \in \mathcal{Y}} P(Y = y) \times P(X = x | Y = y) \tag{19}$$

其中 $P(Y = y)$ 叫做 class prior，$P(X = x|Y = y)$ 叫做 data likelihood given class.