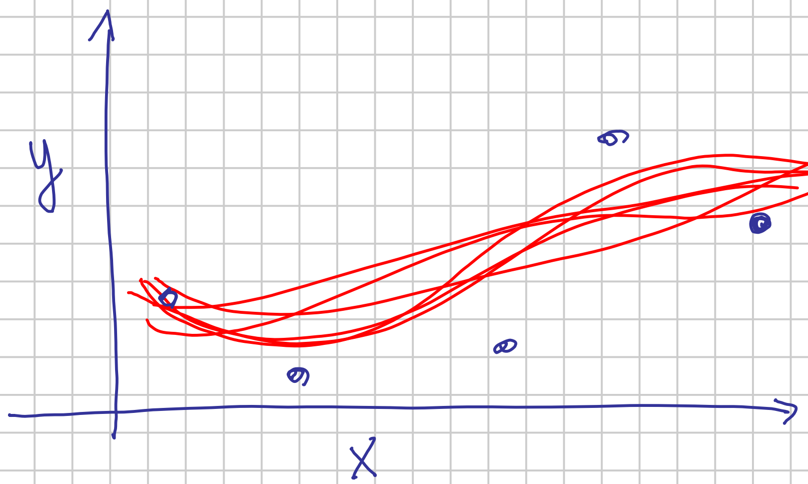
Mathematics for Inference and Machine Learning

Note Title 11/10/2016

Regression



$$y = f(x) + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Function f (what we are looking for) parametrized by θ

example: $y = \underbrace{x^T\theta}_{f(x,\theta)} + \varepsilon$

$y = \underbrace{\phi^T(x)\,\theta}_{f(x,\theta)} + \varepsilon$

Linear models

$\phi(x)$: nonlinear transformation of x

$$f(x, \theta) = \theta_0 + \theta_1 x + \theta_2 x^2$$

Probabilities

$p: \mathbb{R}^D \longrightarrow \mathbb{R}$ probability density function

(i) $\forall x \in \mathbb{R}^D, \quad p(x) \geq 0$

(ii) $\int_{\mathbb{R}^D} p(x)dx = 1$

X: Random variable

Fundamental Rules

marginalization property

(i) Sum Rule $\qquad \int p(x,y)\,dy = p(x)$

(ii) Product Rule $\qquad \underbrace{p(x,y)}_{\text{joint prob distr.}} = p(x)\underbrace{p(y|x)}_{\text{cond. prob. distr.}}$

Notation

x, y random variables

- $p(x,y)$ joint probability distribution
- $p(x|y)$, $p(y|x)$ conditionals
- $p(x)$, $p(y)$ marginals

Bayes' Theorem

prior $\qquad$ likelihood

$$p(x|y) = \frac{p(x)\,p(y|x)}{p(y)}$$

posterior $\qquad$ marginal likelihood / evidence

$$p(y) = \int p(x)\,p(y|x)\,dx$$

Means / Covariances

$x \in \mathbb{R}^D$ random variable, $\qquad x \sim p(x)$

$$E[x] = \int x\, p(x)\,dx \quad \in \mathbb{R}^D$$

$$= \begin{bmatrix} E[x_1] \\ \vdots \\ E[x_D] \end{bmatrix}$$

$x \in \mathbb{R}^D, \; y \in \mathbb{R}^E$

$$\mathrm{cov}[x,y] = E[xy^T] - E[x]E[y]^T \qquad \in \mathbb{R}^{D \times E}$$

$$= \iint (x - E[x])(y - E[y])^T p(x,y)\,dx\,dy$$

$$= \begin{bmatrix} \mathrm{cov}[x_1, y_1] & \mathrm{cov}[x_1, y_2] & \cdots & \mathrm{cov}[x_1, y_E] \\ \vdots & & & \vdots \\ \mathrm{cov}[x_D, y_1] & \cdots & \cdots & \mathrm{cov}[x_D, y_E] \end{bmatrix}$$

$$V[x] = \mathrm{var}[x] = \mathrm{cov}[x,x]$$

$$= \underbrace{\begin{bmatrix} \mathrm{cov}[x_1, x_1] & \cdots & \mathrm{cov}[x_1, x_D] \\ \vdots & \ddots & \vdots \\ \mathrm{cov}[x_D, x_1] & \cdots & \mathrm{cov}[x_D, x_D] \end{bmatrix}}_{} \in \mathbb{R}^{D \times D}$$

- symmetric, posite definite
- Diagonal entries are the marginal variances of $p(x_i) = \int p(x_1, \dots, x_D)\, dx_{\setminus i}$

Sum of Random Variables

$x, y \in \mathbb{R}^D$

$$E[x \pm y] = E[x] \pm E[y]$$

$$V[x \pm y] = V[x] + V[y] \pm \mathrm{cov}[x,y] \pm \mathrm{cov}[y,x]$$

Affine Transformations

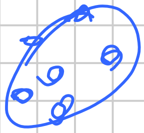
$x \in \mathbb{R}^D$ random variable, $E[x] = \mu$, $V[x] = \Sigma$

$y = Ax + b$ $\qquad$ scalar: $E[ax+b] = a\,E[x] + b$

$$E[y] = E[Ax+b] = A\,E[x] + b = A\mu + b$$

scalar: $var(ax+b) = a^2\, var(x)$

$$V_y[y] = V_x[Ax+b]$$
$$= V[Ax]$$
$$= A\,V[x]\,A^T = A \Sigma A^T$$



$y = Ax+b$

$$cov[x, y] = E[xy^T] - E[x]\,E[y]^T \qquad = \Sigma A^T$$

$$\overset{y=Ax+b}{=} E[x(Ax+b)^T] - E[x]\,E[Ax+b]^T$$

$$= E[x]b^T + E[xx^T]A^T - \mu(A\mu+b)^T$$

$$= \cancel{\mu b^T} + E[xx^T]A^T - \underbrace{\mu\mu^T}_{=E[x]E[x]^T}A^T - \cancel{\mu b^T}$$

$$= \underbrace{\left(E[xx^T] - E[x]\,E[x]^T\right)}_{=\Sigma} A^T$$

$y = f(x)$

$E[y] = \int f(x)\,p(x)\,dx$

$$E[x(Ax+b)^T] = E[x(x^TA^T + b^T)] = E[xx^TA^T] + E[xb^T]$$
$$= E[xx^T]A^T + E[x]b^T$$

Statistical Independence

x, y random variables are stat. independent

$:\Longleftrightarrow \quad p(x,y) = p(x)\,p(y)$

Properties (if x, y independent):

- $p(y|x) = p(y)$
- $p(x|y) = p(x)$
- $V[x+y] = V[x] + V[y]$ $\qquad (cov[x,y] = 0)$

Conditional Independence

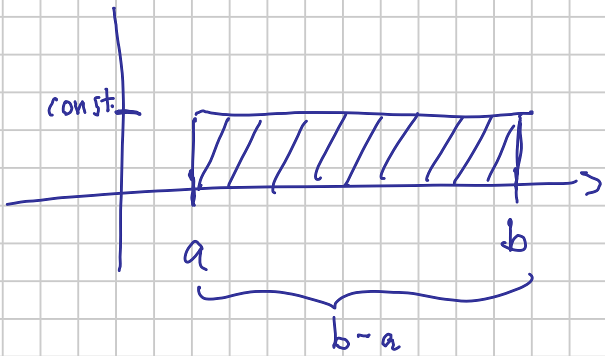
x, y conditionally indep. given z $:\Longleftrightarrow$ $p(x,y \mid z) = p(x \mid z)\, p(y \mid z)$

notation: $x \perp\!\!\!\perp y \mid z$

Probability Distributions

Uniform

$$p(x) = \mathcal{U}[a,b]$$
$$= \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & \text{otherwise} \end{cases}$$



Bernoulli (Binary variables)

- Distribution on a single binary r.v. $x \in \{0,1\}$
- Governed by $\mu \in [0,1]$ that represents the prob. of $x = 1$

$$p(x \mid \mu) = \mu^{x} (1-\mu)^{1-x}, \quad x \in \{0,1\}$$

Binomial (Positive Integers)

- "Repeated Bernoulli experiment" (N repetitions)

$$p(m \mid N, \mu) = \binom{N}{m} \mu^{m} (1-\mu)^{N-m}$$

$\hookrightarrow$ probability of observing m times $x=1$ in a repeated Bernoulli experiment (N times) with $p(x=1) = \mu$

Beta Distribution (Distr. on $[0,1]$)

$\mu \in [0,1]$ random variable

$\alpha, \beta > 0$ just normalization constant

$$p(\mu \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1} (1-\mu)^{\beta-1}$$

Gamma Distribution (positive real numbers)

- $\tau > 0$ random variable
- $a, b > 0$

$$p(\tau \mid a, b) = \mathrm{Gamma}(\tau \mid a, b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} \exp(-b\tau)$$