Machine Learning Notebook

Cong Bao

Contents

Chapter 1

Introduction

1.1 About this Notebook

1.2 Policy of Use

Chapter 2

Mathematics Basics

2.1 Probability

2.1.1 Basic Rules

Three Axioms of Probability Let Ω be a sample space. A probability assigns a real number $P(X)$ to each event $X \subseteq \Omega$ in such a way that

1. $P(X) \geq 0, \forall X$
2. If $X_1, X_2, \ldots$ are pairwise disjoint events ($X_1 \cap X_2 = \emptyset,\ i \neq j,\ i, j = 1, 2, \ldots$), then $P\left(\bigcup_{i=1}^{\infty} X_i\right) = \sum_{i=1}^{\infty} P(X_i)$. (This property is called countable additivity.)
3. $P(\Omega) = 1$

Joint Probability The probability both event A and B occur. $P(X, Y) = P(X \cap Y)$.

Marginalization The probability distribution of any variable in a joint distribution can be recovered by integrating (or summing) over the other variables.

1. For continuous r.v. $P(x) = \int P(x, y)\, dy$; $P(y) = \int P(x, y)\, dx$.
2. For discrete r.v. $P(x) = \sum_y P(x, y)$; $P(y) = \sum_x P(x, y)$.
3. For mixed r.v. $P(x, y) = \sum_w \int P(w, x, y, z)\, dz$, where w is discrete and z is continuous.

Conditional Probibility $P(X = x | Y = y)$ is the probability $X = x$ occurs given the knowledge $Y = y$ occurs. Conditional probability can be extracted from joint probability that

$$P(x | y = y^*) = \frac{P(x, y = y^*)}{\int P(x, y = y^*)\, dx} = \frac{P(x, y = y^*)}{P(y = y^*)}$$

Usually, the formula is written as $P(x|y) = \frac{P(x,y)}{P(y)}$.

Product Rule The formula can be rearranged as $P(x,y) = P(x|y)P(y) = P(y|x)P(x)$. In case of multiple variables

$$\begin{aligned} P(w,x,y,z) &= P(w,x,y|z)P(z) \\ &= P(w,x|y,z)P(y|z)P(z) \\ &= P(w|x,y,z)P(x|y,z)P(y|z)P(z) \end{aligned}$$

Independence If two variables x and y are independent, then r.v. x tells nothing about r.v. y (and vice-versa)

$$\begin{aligned} P(x|y) &= P(x) \\ P(y|x) &= P(y) \\ P(x,y) &= P(x)P(y) \end{aligned}$$

Baye's Rule By rearranging formula in Product Rule, we have

$$\begin{aligned} P(y|x) &= \frac{P(x|y)P(y)}{P(x)} \\ &= \frac{P(x|y)P(y)}{\int P(x,y)\,dy} \\ &= \frac{P(x|y)P(y)}{\int P(x|y)P(y)\,dy} \end{aligned}$$

Expectation Expectation tells us the excepted or average value of some function $f(x)$, taking into account the distribution of x.

$$\begin{aligned} \mathbf{E}[f(x)] &= \sum_x f(x)P(x) \\ \mathbf{E}[f(x)] &= \int f(x)P(x)\,dx \end{aligned}$$

Definition in two dimensions: $\mathbf{E}[f(x,y)] = \iint f(x,y)P(x,y)\,dx\,dy$

Function $f(\bullet)$	Expectation
x^k	k^{th} moment about zero
$(x-\mu_x)^k$	k^{th} moment about the mean

Function $f(\bullet)$	Expectation
x	mean, μ_x
$(x-\mu_x)^2$	variance
$(x-\mu_x)^3$	skew
$(x-\mu_x)^4$	kurtosis
$(x-\mu_x)(x-\mu_y)$	covariance of x and y

Besides, Expectation has the following four rules

1. Expected value of a constant is the constant $\mathbf{E}[\kappa] = \kappa$.
2. Expected value of constant times function is constant times excepted value of function $\mathbf{E}[kf(x)] = k\mathbf{E}[f(x)]$.
3. Expectation of sum of functions is sum of expectation of functions $\mathbf{E}[f(x) + g(y)] = \mathbf{E}[f(x)] + \mathbf{E}[g(x)]$.
4. Expectation of product of functions in variables x and y is product of expectations of functions if x and y are independent $\mathbf{E}[f(x)g(y)] = \mathbf{E}[f(x)]\,\mathbf{E}[g(y)], x \perp\!\!\!\perp y$.

2.1.2 Common Probability Distributions

Bernoulli Bernoulli distribution describes situation where only two possible outcomes $y = 0/y = 1$ or failure/success.

1. $P(x) = \mathbf{Bern}_x[\lambda] = \lambda^x(1-\lambda)^{1-x}$
2. univariate, discrete, binary
3. $x \in \{0, 1\}$; $\lambda \in [0, 1]$
4. $\mathbf{E}[x] = \lambda$, $\mathbf{Var}[x] = \lambda(1-\lambda)$

Beta Beta distribution is the conjugate distribution to Bernoulli distribution.

1. $P(\lambda) = \mathbf{Beta}_\lambda[\alpha, \beta] = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\lambda^{\alpha-1}(1-\lambda)^{\beta-1}$
2. univariate, continuous, unbounded
3. $\lambda \in \mathbb{R}$; $\alpha \in \mathbb{R}^+$, $\beta \in \mathbb{R}^+$
4. $\mathbf{E}[\lambda] = \frac{\alpha}{\alpha+\beta}$, $\mathbf{Var}[\lambda] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Categorical Categorical distribution describes situation with K possible outcomes.

1. $P(x) = \mathbf{Cat}_x[\boldsymbol{\lambda}]$, $P(x = k) = \lambda_k$, $P(\boldsymbol{x} = \boldsymbol{e}_k) = \prod_{j=1}^{K} \lambda_j^{x_j} = \lambda_k$
2. univariable, discrete, multi-valued
3. $x \in \{1, 2, \ldots, K\}$; $\lambda_k \in [0, 1]$ where $\sum_k \lambda_k = 1$
4. $\mathbf{E}[x_i] = \lambda_i$, $\mathbf{Var}[x_i] = \lambda_i(1-\lambda_i)$, $\mathbf{Cov}[x_i, x_j] = -\lambda_i\lambda_j$ $(i \neq j)$

Dirichlet Dirichlet distribution is the conjugate distribution to categorical distribution.

1. $P(\boldsymbol{\lambda}) = \mathbf{Dir}_{\boldsymbol{\lambda}}[\boldsymbol{\alpha}] = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\lambda_k^{\alpha_k-1}$
2. multivariable, continuous, bounded, sums to one
3. $\boldsymbol{x} = [x_1, x_2, \ldots, x_K]^\top$, $x_k \in [0, 1]$, $\sum_{k=1}^{K} x_k = 1$; $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_K]$, $\alpha_k \in \mathbb{R}^+$
4. $\mathbf{E}[\lambda_i] = \frac{\alpha_i}{\sum_k \alpha_k}$, $\mathbf{Var}[\lambda_i] = \frac{\alpha_i(\sum_k \alpha_k - \alpha_i)}{(\sum_k \alpha_k)^2(\sum_k \alpha_k+1)}$, $\mathbf{Cov}[\lambda_i, \lambda_j] = \frac{-\alpha_i\alpha_j}{(\sum_k \alpha_k)^2(\sum_k \alpha_k+1)}$ $(i \neq j)$

Univariable Normal Univariable normal distribution describes single continuous variable.

1. $P\left(x\right) = \mathbf{Norm}_x[\mu, \sigma^2] = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
2. univariable, continuous, unbounded
3. $x \in \mathbb{R}$; $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+$
4. $\mathbf{E}\left[x\right] = \mu$, $\mathbf{Var}\left[x\right] = \sigma^2$

Normal Inverse Gamma

Multivariate Normal

Normal inverse Wishart

2.2 Linear Algebra

2.3 Calculus

2.4 Informatics

2.5 Optimization

Chapter 3

Machine Learning Basics

Chapter 4

Regression

4.1 Linear Regression

4.2 Non-linear Regression

4.3 Logistic Regression

Chapter 5

Support Vector Machines

Bibliography

Appendix A

Test

test