

# Machine Learning Notebook

Cong Bao

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	About this Notebook . . . . .	1
1.2	Policy of Use . . . . .	1
<b>2</b>	<b>Mathematics Basics</b>	<b>2</b>
2.1	Probability . . . . .	2
2.1.1	Basic Rules . . . . .	2
2.1.2	Common Probability Distributions . . . . .	4
2.2	Linear Algebra . . . . .	6
2.3	Calculus . . . . .	6
2.4	Informatics . . . . .	6
2.5	Optimization . . . . .	6
<b>3</b>	<b>Machine Learning Basics</b>	<b>7</b>
3.1	Regularization . . . . .	7
3.1.1	Under-fitting & Over-fitting . . . . .	7
3.1.2	Bias & Variance . . . . .	7
3.1.3	Vector Norm . . . . .	7
3.1.4	Penalize Complexity . . . . .	8
3.2	Cross-Validation . . . . .	8
3.3	Bayesian Learning . . . . .	8
3.3.1	Bayes' Rule Terminology . . . . .	8
3.3.2	Maximum Likelihood . . . . .	9
3.3.3	Maximum a Posterior (MAP) . . . . .	9
3.3.4	Bayesian Approach . . . . .	9
3.3.5	Example: Normal Distribution . . . . .	10
3.3.6	Example: Categorical Distribution . . . . .	10
3.4	Machine Learning Models . . . . .	10

## CONTENTS

---

<b>4</b>	<b>Regression</b>	<b>11</b>
4.1	Linear Regression . . . . .	11
4.2	Non-linear Regression . . . . .	11
4.3	Logistic Regression . . . . .	11
<b>5</b>	<b>Support Vector Machines</b>	<b>12</b>
	<b>Bibliography</b>	<b>13</b>
<b>A</b>	<b>Test</b>	<b>14</b>

# 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  $\Omega$  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  $X_1, X_2, \dots$  are pairwise disjoint events ( $X_1 \cap X_2 = \emptyset, i \neq j, i, j = 1, 2, \dots$ ), then  $P(\bigcup_{i=1}^{\infty} X_i) = \sum_{i=1}^{\infty} P(X_i)$ . (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 Probability**  $P(X = x|Y = y)$  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 expected 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
$x$	mean, $\mu_x$
$(x - \mu_x)^2$	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 expected value of function  $\mathbf{E}[kf(x)] = k\mathbf{E}[f(x)]$ .
3. Expectation of sum of functions is sum of expectation of functions  $\mathbf{E}[f(x) + g(y)] = \mathbf{E}[f(x)] + \mathbf{E}[g(y)]$ .
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.  $P(x) = \mathbf{Cat}_x[\boldsymbol{\lambda}]$ ,  $P(x = k) = \lambda_k$ ,  $P(\mathbf{x} = \mathbf{e}_k) = \prod_{j=1}^K \lambda_j^{x_j} = \lambda_k$
2. univariate, discrete, multi-valued
3.  $x \in \{1, 2, \dots, K\}$ ;  $\lambda_k \in [0, 1]$  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.  $P(\boldsymbol{\lambda}) = \mathbf{Dir}_\lambda[\boldsymbol{\alpha}] = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \lambda_k^{\alpha_k-1}$
2. multivariate, continuous, bounded, sums to one
3.  $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \dots, \lambda_K]^\top$ ,  $\lambda_k \in [0, 1]$ ,  $\sum_{k=1}^K \lambda_k = 1$ ;  $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_K]$ ,  $\alpha_k \in \mathbb{R}_+$
4.  $\mathbf{E}[\lambda_i] = \frac{\alpha_i}{\sum_k \alpha_k}$ ,  $\mathbf{Var}[\lambda_i] = \frac{\alpha_i(\sum_k \alpha_k - \alpha_i)}{(\sum_k \alpha_k)^2(\sum_k \alpha_k + 1)}$ ,  $\mathbf{Cov}[\lambda_i, \lambda_j] = \frac{-\alpha_i\alpha_j}{(\sum_k \alpha_k)^2(\sum_k \alpha_k + 1)}$  ( $i \neq j$ )

**Univariate Normal** Univariate normal distribution describes single continuous variable.

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

**Normal Inverse Gamma** Normal inverse gamma distribution is a conjugate distribution to univariate normal distribution.

1.  $P(\mu, \sigma^2) = \mathbf{NormInvGam}_{\mu, \sigma^2}[\alpha, \beta, \gamma, \delta] = \frac{\sqrt{\gamma}}{\sqrt{2\pi\sigma^2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left(-\frac{2\beta+\gamma(\delta-\mu)^2}{2\sigma^2}\right)$
2. bivariate, continuous,  $\mu$  unbounded,  $\sigma^2$  bounded below
3.  $\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+; \alpha \in \mathbb{R}_+, \beta \in \mathbb{R}_+, \gamma \in \mathbb{R}_+, \delta \in \mathbb{R}$
4.  $\mathbf{E}[\mu] = \delta, \mathbf{E}[\sigma^2] = \frac{\beta}{\alpha-1} \ (\alpha > 1), \mathbf{Var}[\mu] = \frac{\beta}{(\alpha-1)\gamma} \ (\alpha > 1), \mathbf{Var}[\sigma^2] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} \ (\alpha > 2), \mathbf{Cov}[\mu, \sigma^2] = 0 \ (\alpha > 1)$

**Multivariate Normal** Multivariate normal distribution describes multiple continuous variables. It takes two parameters: a vector containing mean position  $\boldsymbol{\mu}$ , and a symmetric positive definite covariance matrix  $\boldsymbol{\Sigma}$ .

1.  $P(\mathbf{x}) = \mathbf{Norm}_x[\boldsymbol{\mu}, \boldsymbol{\Sigma}] = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$
2. multivariate, continuous, unbounded
3.  $\mathbf{x} \in \mathbb{R}^K; \boldsymbol{\mu} \in \mathbb{R}^K, \boldsymbol{\Sigma} \in \mathbb{R}^{K \times K}$  (positive semi-definite matrix)
4.  $\mathbf{E}[\mathbf{x}] = \boldsymbol{\mu}, \mathbf{Var}[\mathbf{x}] = \boldsymbol{\Sigma}$

**Normal Inverse Wishart** Normal inverse wishart distribution is a conjugate distribution to multivariate normal distribution.

1.  $P(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathbf{NormInvWis}_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}[\alpha, \boldsymbol{\Psi}, \gamma, \boldsymbol{\delta}]$   
 $= \frac{\gamma^{D/2} |\boldsymbol{\Psi}|^{\alpha/2} |\boldsymbol{\Sigma}|^{-\frac{\alpha+D+2}{2}}}{(2\pi)^{D/2} 2^{(\alpha\boldsymbol{\Sigma})/2} \Gamma_D(\alpha/2)} \exp\left(-\frac{1}{2}(\text{Tr}(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}) + \gamma(\boldsymbol{\mu} - \boldsymbol{\delta})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\delta}))\right)$
2. multivariate,  $\boldsymbol{\mu}$  unbounded,  $\boldsymbol{\Sigma}$  square, positive definite
3.  $\boldsymbol{\mu} \in \mathbb{R}^K, \boldsymbol{\Sigma} \in \mathbb{R}^{K \times K}; \alpha \in \mathbb{R}_{>D-1}, \boldsymbol{\Psi} \in \mathbb{R}^{K \times K}, \gamma \in \mathbb{R}_+, \boldsymbol{\delta} \in \mathbb{R}^K$



## **2.2 Linear Algebra**

## **2.3 Calculus**

## **2.4 Informatics**

## **2.5 Optimization**

# Chapter 3

## Machine Learning Basics

### 3.1 Regularization

#### 3.1.1 Under-fitting & Over-fitting

**Under-fitting** If  $N > D$  (e.g. 30 data points, 2 dimensions) we have more equations than unknowns: over-determined system. Input-output relations can only hold approximately.

**Over-fitting** If  $N < D$  (e.g. 30 points, 15265 dimensions) we have more unknowns than equations: under-determined system. Input-output equations hold exactly, but we are simply memorizing data.

#### 3.1.2 Bias & Variance

**High Bias & Low Variance** A rigid model's (low complexity) performance is more predictable in the test set but the model may not be good even on the training set.

**Low Bias & High Variance** A flexible model (high complexity) approximates the target function well in the training set but can "overtrain" and have poor performance on the test set.

#### 3.1.3 Vector Norm

**L1, ("Manhattan") norm**  $\|\mathbf{w}\|_1 = \sum_{d=1}^D |w_d|$

**L2, ("Euclidean") norm**  $\|\mathbf{w}\|_2 = \sqrt{\sum_{d=1}^D w_d^2} = \sqrt{\langle \mathbf{w}, \mathbf{w} \rangle} = \sqrt{\mathbf{w}^\top \mathbf{w}}$

**Lp norm,  $p > 1$**   $\|\mathbf{w}\|_p = \left( \sum_{d=1}^D w_d^p \right)^{1/p}$

### 3.1.4 Penalize Complexity

In linear regression, the residual vector is  $\epsilon = \mathbf{y} - \Psi\mathbf{w}$ . The loss function is  $L(\mathbf{w}) = \epsilon^\top \epsilon$ . We add a complexity term  $R(\mathbf{w}) = \|\mathbf{w}\|_2^2 = \mathbf{w}^\top \mathbf{w}$  to the loss function. Hence, the original loss function becomes  $L(\mathbf{w}) = \epsilon^\top \epsilon + \lambda \mathbf{w}^\top \mathbf{w}$ .

Without regularization, the loss function is  $L(\mathbf{w}) = \epsilon^\top \epsilon$ . Let  $\nabla L(\mathbf{w}^*) = 0$ , we have  $\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ .

With L2-regularization, the loss function is  $L(\mathbf{w}) = \epsilon^\top \epsilon + \lambda \mathbf{w}^\top \mathbf{w}$ . Let  $\nabla L(\mathbf{w}^*) = 0$ , we have  $\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$ . The additional  $\lambda \mathbf{I}$  makes the data matrix more robust to calculate inversion.

## 3.2 Cross-Validation

We can select hyperparameters with (cross-)validation. Cross-validation excludes part of the training data from parameter estimation, and use them only to predict the test error.

K-fold cross validation: split data set into K folds and each time train on (K-1) folds and valid on the remaining fold until all folds have been used as validation fold. The cross-validation error is the average of K validation errors. We pick hyperparameters that minimize cross-validation error.

## 3.3 Bayesian Learning

### 3.3.1 Bayes' Rule Terminology

Bayes' Rule:

$$P(y|x) = \frac{P(x|y) P(y)}{\int P(x|y) P(y) dy}$$

**Prior**  $P(y)$  what we know about  $y$  before seeing  $x$ . In parameters learning we choose prior that is conjugate to likelihood.

**Likelihood**  $P(x|y)$  propensity for observing a certain value of  $x$  given a certain value of  $y$ .

**Posterior**  $P(y|x)$  what we know about  $y$  after seeing  $x$ . Posterior must have same form as conjugate prior distribution.

**Evidence**  $\int P(x|y) P(y) dy$  a constant to ensure that the LHS is a valid distribution. Posterior must be a distribution which implies that evidence equals to a constant  $\kappa$  from conjugate relation.

### 3.3.2 Maximum Likelihood

**Fitting** As the name suggests we find the parameters under which the data  $\mathbf{x}_{1...I}$  are most likely. Here, we have assumed that data was independent.

$$\begin{aligned}\hat{\theta} &= \operatorname{argmax}_{\theta} P(\mathbf{x}_{1...I}|\theta) \\ &= \operatorname{argmax}_{\theta} \prod_{i=1}^I P(\mathbf{x}_i|\theta)\end{aligned}$$

**Predictive Density** Evaluate new data point  $\mathbf{x}^*$  under probability distribution  $P(\mathbf{x}^*|\hat{\theta})$  with best parameters.

### 3.3.3 Maximum a Posterior (MAP)

**Fitting** As the name suggests we find the parameters which maximize the posterior probability  $P(\theta|\mathbf{x}_{1...I})$ . Again we have assumed that data was independent.

$$\begin{aligned}\hat{\theta} &= \operatorname{argmax}_{\theta} P(\theta|\mathbf{x}_{1...I}) \\ &= \operatorname{argmax}_{\theta} \frac{P(\mathbf{x}_{1...I}|\theta) P(\theta)}{P(\mathbf{x}_{1...I})} \\ &= \operatorname{argmax}_{\theta} \frac{\prod_{i=1}^I P(\mathbf{x}_i|\theta) P(\theta)}{P(\mathbf{x}_{1...I})}\end{aligned}$$

Since the denominator does not depend on the parameters we can instead maximize

$$\hat{\theta} = \operatorname{argmax}_{\theta} \prod_{i=1}^I P(\mathbf{x}_i|\theta) P(\theta)$$

**Predictive Density** Evaluate new data point  $\mathbf{x}^*$  under probability distribution with MAP parameters  $P(\mathbf{x}^*|\hat{\theta})$

### 3.3.4 Bayesian Approach

**Fitting** Compute the posterior distribution over possible parameter values using Bayes' rule. Principle: There are many values that could have explained the data. Instead of picking one set of parameters, try to capture all of the possibilities.

$$P(\theta|\mathbf{x}_{1...I}) = \frac{\prod_{i=1}^I P(\mathbf{x}_i|\theta) P(\theta)}{P(\mathbf{x}_{1...I})}$$

**Predictive Density** (a) Each possible parameter value makes a prediction. (b) Some parameters more probable than others.

$$P(\mathbf{x}^*|\mathbf{x}_{1...I}) = \int P(\mathbf{x}^*|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathbf{x}_{1...I}) d\boldsymbol{\theta}$$

Make a prediction that is an infinite weighted sum (integral) of the predictions for each parameter value ( $P(\mathbf{x}^*|\boldsymbol{\theta})$ ), where weights are the probabilities ( $P(\boldsymbol{\theta}|\mathbf{x}_{1...I})$ ).

### 3.3.5 Example: Normal Distribution

### 3.3.6 Example: Categorical Distribution

## 3.4 Machine Learning Models

# 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