

# Contents

<b>1</b>	<b>Introduction and Probability Theory</b>	<b>5</b>
1.1	Machine Learning Basics . . . . .	5
1.1.1	Terminologies . . . . .	5
1.1.2	Supervised vs Unsupervised Learning . . . . .	5
1.1.3	Probability Theory . . . . .	6
<b>2</b>	<b>Statistical Schools of Thought</b>	<b>7</b>
2.1	Frequentist statistics . . . . .	7
2.2	Bayesian Statistics . . . . .	8
2.3	Parametric vs. None-parametric models . . . . .	9
2.4	Generative vs. Discriminative models . . . . .	9
<b>3</b>	<b>Linear Regression and Optimization</b>	<b>10</b>
3.1	Linear Regression via Decision Theory . . . . .	10
3.2	Linear Regression via Frequentist Probabilistic Model . . . . .	10
3.3	Non-Linear continuous Optimization . . . . .	11
3.4	Affine and Convex Sets . . . . .	12
3.4.1	Excursion: (Semi) Definite Matrices . . . . .	13
3.4.2	Convex Sets (Continued) . . . . .	13
3.5	Convex Function . . . . .	14
3.6	Convex Optimization . . . . .	15
3.6.1	Operations That Preserve Convexity . . . . .	16
3.7	Standard Form of Convex Optimization Problems . . . . .	16
3.8	Optimality Condition for Convex Problems . . . . .	17
3.8.1	L1 and L2 Norm . . . . .	17
<b>4</b>	<b>Logistic Regression and Basis Expansion</b>	<b>19</b>
4.1	Logistic Regression: Binary Classification . . . . .	19
4.2	Logistic Regression: Decision-Theoretical View . . . . .	19
4.3	Basis Expansion . . . . .	20
<b>5</b>	<b>Regularization</b>	<b>21</b>
5.1	Regularization in Linear Models . . . . .	21
5.1.1	Co-linearity . . . . .	21
5.1.2	Regulariser as a prior . . . . .	22
5.2	Regularization in Non-linear Models . . . . .	22
5.3	Regularization as a constraint . . . . .	23
5.4	Bias-Variance Tradeoff . . . . .	23
<b>6</b>	<b>Perceptron</b>	<b>24</b>
<b>7</b>	<b>Multilayer Perceptron and Backpropagation</b>	<b>26</b>
7.1	Multi-Layer Perceptron . . . . .	26
<b>8</b>	<b>Deep Learning, Convolutional ANNs and Autoencoders</b>	<b>28</b>
8.1	Deep Learning and Representation Learning . . . . .	28
8.2	Convolutionary Neural Networks (CNN) . . . . .	29
8.3	Auto-encoder . . . . .	29

<b>9</b>	<b>Support Vector Machines</b>	<b>30</b>
9.1	Maximum-Margin Classifier: Hard Margin SVM . . . . .	30
<b>10</b>	<b>Soft-Margin SVM, Lagrangian Duality</b>	<b>33</b>
10.1	Soft-Margin SVMs . . . . .	33
10.2	Lagrangian Duality for SVM . . . . .	34
<b>11</b>	<b>Kernel Methods</b>	<b>36</b>
11.1	Kernelising the SVM . . . . .	36
11.2	Modular Learning . . . . .	37
11.3	Constructing Kernels . . . . .	38
<b>12</b>	<b>Ensemble Methods</b>	<b>40</b>
12.1	Bagging . . . . .	40
12.2	Boosting . . . . .	41
12.3	Stacking . . . . .	43
<b>13</b>	<b>Multi-Armed Bandits (Assignment 2)</b>	<b>44</b>
13.1	Stochastic MAB Setting . . . . .	44
13.2	e-Greedy . . . . .	44
13.2.1	Pure Greedy vs. e-greedy . . . . .	44
13.2.2	Q initial values: Optimistic vs. Pessimism . . . . .	44
13.3	Limitation of e-greedy . . . . .	45
13.4	Upper Confidence Bound (UCB) Algorithm . . . . .	45
13.4.1	UCB vs. e-greedy . . . . .	45
<b>14</b>	<b>Gaussian Mixture Model, Expectation Maximization</b>	<b>46</b>
14.1	Unsupervised Learning . . . . .	46
14.2	Gaussian Mixture Model . . . . .	46
14.3	Expectation Maximization Algorithm . . . . .	47
14.3.1	Motivation of EM . . . . .	48
14.3.2	Resolve the MLE Puzzle: Introduce latent variable . . . . .	48
14.4	EM for GMM . . . . .	49
14.4.1	E-Step for GMM . . . . .	50
14.4.2	M-Step for GMM . . . . .	50
14.4.3	K-means as a EM model for restricted GMM model . . . . .	51
14.4.4	How to choose $k$ . . . . .	51
<b>15</b>	<b>Dimensionality Reduction: PCA</b>	<b>52</b>
15.1	Principle Component Analysis (PCA) . . . . .	52
15.2	Non-linear data and Kernel PCA . . . . .	54
<b>16</b>	<b>Bayesian Regression</b>	<b>55</b>
16.1	Problem with Frequentists: Uncertainty . . . . .	55
16.2	Bayesian view on Uncertainty . . . . .	55
16.3	Bayesian Regression . . . . .	56
16.4	Bayesian Prediction . . . . .	57
<b>17</b>	<b>Bayesian Classification</b>	<b>59</b>
17.1	Beta-Binomial Conjugacy . . . . .	59
17.2	Bayesian Logistic Regression . . . . .	59

<b>18 Probabilistic Graphical Models (PGM)</b>	<b>61</b>
18.1 Discrete Joint Distribution Tables . . . . .	61
18.2 Directed PGMs . . . . .	62
18.2.1 Example . . . . .	62
18.2.2 PGM Bayesian or Frequentist? . . . . .	63
18.3 Undirected PGMs: Markov Random Field . . . . .	63
18.4 Example PGMs . . . . .	64
18.4.1 Hidden Markov Model (HMM) AND Kalman Filter . . . . .	64
18.4.2 Conditional Random Fields (CRF) . . . . .	64
<b>19 PGM Inference</b>	<b>65</b>
<b>20 MLE, MAP, Bayesian Regression and Bayesian Inference</b>	<b>67</b>
20.1 Bayes Rule . . . . .	67
20.2 Maximum Likelihood Estimates . . . . .	67
20.3 Maximum a Posterior . . . . .	67
20.4 Bayesian Regression . . . . .	67
20.5 Bayesian Inference (Prediction) . . . . .	68

# 1 Introduction and Probability Theory

An important hypothesis of Machine Learning

Pre-existing data repositories contain a lot of potentially valuable knowledge

Definition: What is Learning

(Semi-) automatic extraction of **valid, novel, useful and comprehensible knowledge** - in the form of rules, regularities, patterns, constraints or models - from arbitrary sets of data.

The goal is to develop efficient and useful algorithm.

**Contents this subject covers:**

Foundations of statistical learning, linear models, non-linear bases, kernel approaches, neural networks, Bayesian learning, probabilistic graphical models (Bayes Nets, Markov Random Fields), cluster analysis, dimensionality reduction, regularization and model selection

## 1.1 Machine Learning Basics

### 1.1.1 Terminologies

- **Instance:** measurements about individual entities/objects: e.g. a loan application (data points)
- **Attribute (Feature, explanatory variable):** component of the instances: e.g. the applicant's salary, numerics, etc. ( $x_i$ )
- **Label (Response, dependent variable):** an outcome that is categorical, numbers, etc. ( $y$ )
- **Examples:** instance coupled with label, i.e.  $\langle x_i, y \rangle$
- **Models:** discovered relationship between attributes and/or label.

### 1.1.2 Supervised vs Unsupervised Learning

	Data	Model used for
Supervised Learning	Labelled	Predict labels on new instances
Unsupervised Learning	Unlabelled	Cluster related instances; Project to fewer dimensions; Understand attribute relationships

**Evaluations: important for supervised learning**

**Evaluation principle:** measure quality is problem-dependent

**Step 1: pick an evaluation metric** Accuracy, Contingency table, Precision-Recall, F1, ROC curves

**step 2:** procure an independent, labelled **test sets**

**Step 3:** Average the evaluation metric over the test sets.

**Cross Validation** especially when data is poor.

### 1.1.3 Probability Theory

#### Three types of models

$\hat{y} = f(x)$ : regressions, best fitting parameters given a model, we model  $x$ .  
 $p(y|x)$ : For a given  $x$ , we model  $y$  as a distribution (likelihood of  $y$  given  $x$ ).  
 $p(x, y)$ : we model  $(x, y)$  together, the probability of having  $(x, y)$ .

#### Definition: Random Variable

A random variable  $X$  is a **numerical function** of outcome  $X(\omega) \in R$

#### Discrete distribution

- Govern r.v. taking discrete values
- Described by **probability mass function**  $p(x)$  which is  $p(X = x)$
- $P(X \leq x) = \sum_{a=-\infty}^x p(a)$
- **Examples:** Bernoulli, Binomial, Multinomial, Poisson

#### Continuous distribution

- Govern real-valued r.v.
- Described by **probability density function**  $p(x)$  which is  $p(X = x)$
- $P(X \leq x) = \int_{-\infty}^x p(a) da$  (Sum from the very left)
- **Examples:** Uniform, Normal, Laplace, Gamma, Beta, Dirichlet

#### Definition: Bayes' Theorem

In terms of events  $A, B$ :

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \Leftrightarrow P(A|B)P(B) = P(B|A)P(A)$$

Bayesian statistical inference makes heavy use of:

- **Marginals:** probabilities of individual variables
- **Marginalisation:** summing away all but r.v.'s of interest (reduce the number of variables/distribution)

## 2 Statistical Schools of Thought

### 2.1 Frequentist statistics

Wherein unknown model parameters are treated as having fixed but unknown values.

The problem that we intend to solve

**Given:**  $X_1, X_2, \dots, X_n$  drawn i.i.d from some unknown distribution

**Want to:** identify unknown distribution

**Approach:** Parametric Approach

- **Some model:** parameterised by parameter sets  $\theta$
- **Point estimates:** point estimates  $\hat{\theta}$  a function or statistics of data.

Definition: Bias-Variance Decomposition

- **Bias:**  $B_{\theta}(\hat{\theta}) = E_{\theta}[\hat{\theta}(X_1, \dots, X_n)] - \theta$
- **Variance:**  $Var_{\theta}(\hat{\theta}) = E_{\theta}[(\hat{\theta} - E_{\theta}[\hat{\theta}])^2]$
- **Bias-variance decomposition of square loss:**

$$E_{\theta}[(\theta - \hat{\theta})^2] = [B(\theta)]^2 + Var_{\theta}(\hat{\theta})$$

What we really care is the squared loss, which contains both bias and variance. (Notice that empirically, we care about the expected loss, since we don't know the true distribution, we use the distribution of the sample to approximate).

Asymptotic properties

**Consistency:**  $\hat{\theta}(x_1, \dots, x_N)$  converges to true  $\theta$  as  $n \rightarrow \infty$  (i.e., if we increase the sample size to infinity, does the estimated distribution converge to the true distribution?)

**Efficiency:** asymptotic variance is as small as possible.

The Approaches: Maximum Likelihood Estimation (MLE)

---

**Algorithm 1** Maximum Likelihood Estimation (MLE)

---

- 1: We have a set of data  $(X_1, \dots, X_n)$
- 2: We propose a distribution,  $p_\theta$ , which is assumed to have generated the data (i.e. we assume that the data is generated by using this distribution function)
- 3: Express the likelihood of the data:

$$\prod_{i=1}^n p_\theta(X_i)$$

- 4: Apply log trick
  - 5: Optimise to find the best (most likely) parameters  $\theta$ : two ways to do that
  - 6: (1) F.O.C: take partial derivatives of log likelihood w.r.t  $\hat{\theta}$
  - 7: (2) iterative processes, Newton method etc.
- 

## 2.2 Bayesian Statistics

Wherein unknown model parameters have associated distributions reflecting prior belief.

Key idea: **Probabilities**  $\Leftrightarrow$  **beliefs**

### Bayesian Machine Learning:

**Step 1:** Start with prior  $P(\theta)$  and likelihood  $P(X|\theta)$

**step 2:** Observe data  $X = x$

**Step 3:** Update prior to posterior  $P(\theta|X = x)$

#### Definition: Bayes' Theorem

In terms of events A, B:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \Leftrightarrow P(A|B)P(B) = P(B|A)P(A)$$

Bayesian statistical inference makes heavy use of:

- **Marginals:** probabilities of individual variables
- **Marginalisation:** summing away all but r.v.'s of interest (reduce the number of unwanted variables/distribution)

$$P(X = x) = \sum_t P(X = x, \theta = t)$$