https://github.com/soulmachine/machine-learning-cheat-sheet
soulmachine@gmail.com

Machine Learning Cheat Sheet

Classical equations and diagrams in machine learning

May 16, 2013

Preface

This cheat sheet contains many classical equations and diagrams on machine learning, which will help you quickly recall knowledge and ideas on machine learning.

The cheat sheet will also appeal to someone who is preparing for a job interview related to machine learning.

At Tsinghua University, May 2013 *soulmachine*

Contents

Notation

Introduction

It is very difficult to come up with a single, consistent notation to cover the wide variety of data, models and algorithms that we discuss. Furthermore, conventions difer between machine learning and statistics, and between different books and papers. Nevertheless, we have tried to be as consistent as possible. Below we summarize most of the notation used in this book, although individual sections may introduce new notation. Note also that the same symbol may have diferent meanings depending on the context, although we try to avoid this where possible.

General math notation

Symbol	Meaning
$\lfloor x \rfloor$	Floor of x, i.e., round down to nearest integer
$\lceil x \rceil$	Ceiling of x, i.e., round down to nearest integer
$\mathbf{x} \otimes \mathbf{y}$	Convolution of **x** and **y**
$\mathbf{x} \odot \mathbf{y}$	Hadamard (elementwise) product of **x** and **y**
$a \wedge b$	logical AND
$a \vee b$	logical OR
$\neg a$	logical NOT
$\mathbb{I}(x)$	Indicator function, $\mathbb{I}(x) = 1$ if x is true, else $\mathbb{I}(x) = 0$
∞	Infinity
$\rightarrow$	Tends towards, e.g., $n \rightarrow \infty$
$\propto$	Proportional to, so $y = ax$ can be written as $y \propto x$
$\lvert x \rvert$	Absolute value
$\lvert \mathcal{S} \rvert$	Size (cardinality) of a set
$n!$	Factorial function

∇	Vector of first derivatives
∇^2	Hessian matrix of second derivatives
$\triangleq$	Defined as
$O(\cdot)$	Big-O: roughly means order of magnitude
$\mathbb{R}$	The real numbers
$1:n$	Range (Matlab convention): $1:n = 1,2,...,n$
$\approx$	Approximately equal to
$\arg\max_x f(x)$	Argmax: the value x that maximizes f
$B(a,b)$	Beta function, $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$
$B(\alpha)$	Multivariate beta function, $\frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}$
$\binom{n}{k}$	n choose k , equal to $n!/(k!(nk)!)$
$\delta(x)$	Dirac delta function,$\delta(x) = \infty$ if $x = 0$, else $\delta(x) = 0$
$exp(x)$	Exponential function e^x
$\Gamma(x)$	Gamma function, $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$
$\Psi(x)$	Digamma function,$Psi(x) = \frac{d}{dx}\log\Gamma(x)$
$\mathcal{X}$	A set from which values are drawn (e.g.,$\mathcal{X} = \mathbb{R}^D$)

Linear algebra notation

We use boldface lower-case to denote vectors, such as **x**, and boldface upper-case to denote matrices, such as **X**. We denote entries in a matrix by non-bold upper case letters, such as X_{ij}. Vectors are assumed to be column vectors, unless noted otherwise.

Symbol	Meaning
$\mathbf{X} \succ 0$	**X** is a positive definite matrix
$tr(\mathbf{X})$	Trace of a matrix
$det(\mathbf{X})$	Determinant of matrix **X**
$\lvert\mathbf{X}\rvert$	Determinant of matrix **X**
$\mathbf{X}^{-1}$	Inverse of a matrix
$\mathbf{X}^{\dagger}$	Pseudo-inverse of a matrix
$\mathbf{X}^T$	Transpose of a matrix
$\mathbf{x}^T$	Transpose of a vector
$diag(x)$	Diagonal matrix made from vector **x**
$diag(X)$	Diagonal vector extracted from matrix **X**
$\mathbf{I}$ or $\mathbf{I}_d$	Identity matrix of size $d \times d$ (ones on diagonal, zeros of)
$\mathbf{1}$ or $\mathbf{1}_d$	Vector of ones (of length d)
$\mathbf{0}$ or $\mathbf{0}_d$	Vector of zeros (of length d)

$\|\|\mathbf{x}\|\| = \|\|\mathbf{x}\|\|_2$	Euclidean or ℓ_2 norm $\sqrt{\sum_{j=1}^{d} x_j^2}$
$\|\|\mathbf{x}\|\|_1$	ℓ_1 norm $\sum_{j=1}^{d} \left\|x_j\right\|$
$\mathbf{X}_{:,j}$	jth column of matrix
$\mathbf{X}_{i,:}$	transpose of ith row of matrix (a column vector)
$\mathbf{X}_{i,j}$	Element (i,j) of matrix $\mathbf{X}$
$\mathbf{x} \otimes \mathbf{y}$	Tensor product of $\mathbf{x}$ and $\mathbf{y}$

Probability notation

We denote random and fixed scalars by lower case, random and fixed vectors by bold lower case, and random and fixed matrices by bold upper case. Occasionally we use non-bold upper case to denote scalar random variables. Also, we use $p()$ for both discrete and continuous random variables

Symbol	Meaning
X,Y	Random variable
$P()$	Probability of a random event
$F()$	Cumulative distribution function(CDF), also called distribution function
$p(x)$	Probability mass function(PMF)
$f(x)$	probability density function(PDF)
$F(x,y)$	Joint CDF
$p(x,y)$	Joint PMF
$f(x,y)$	Joint PDF
$p(X\|Y)$	Conditional PMF, also called conditional probability
$f_{X\|Y}(x\|y)$	Conditional PDF
$X \perp Y$	X is independent of Y
$X \not\perp Y$	X is not independent of Y
$X \perp Y\|Z$	X is conditionally independent of Y given Z
$X \not\perp Y\|Z$	X is not conditionally independent of Y given Z
$X \sim p$	X is distributed according to distribution p
$\boldsymbol{\alpha}$	Parameters of a Beta or Dirichlet distribution
$cov[X]$	Covariance of X
$\mathbb{E}[X]$	Expected value of X
$\mathbb{E}_q[X]$	Expected value of X wrt distribution q
$\mathbb{H}(X)$ or $\mathbb{H}(p)$	Entropy of distribution $p(X)$
$\mathbb{I}(X;Y)$	Mutual information between X and Y
$\mathbb{KL}(p\|\|q)$	KL divergence from distribution p to q
$\ell(\boldsymbol{\theta})$	Log-likelihood function
$L(\boldsymbol{\theta},a)$	Loss function for taking action a when true state of nature is $\boldsymbol{\theta}$

λ	Precision (inverse variance) $\lambda = 1/\sigma^2$
Λ	Precision matrix $\Lambda = \Sigma^{-1}$
mode[**X**]	Most probable value of **X**
μ	Mean of a scalar distribution
$\boldsymbol{\mu}$	Mean of a multivariate distribution
Φ	cdf of standard normal
ϕ	pdf of standard normal
π	multinomial parameter vector, Stationary distribution of Markov chain
ρ	Correlation coefficient
sigm(x)	Sigmoid (logistic) function, $\frac{1}{1+e^{-x}}$
σ^2	Variance
Σ	Covariance matrix
var[x]	Variance of x
ν	Degrees of freedom parameter
Z	Normalization constant of a probability distribution

Machine learning/statistics notation

In general, we use upper case letters to denote constants, such as C,K,M,N,T, etc. We use lower case letters as dummy indexes of the appropriate range, such as $c = 1 : C$ to index classes, $i = 1 : M$ to index data cases, $j = 1 : N$ to index input features, $k = 1 : K$ to index states or clusters, $t = 1 : T$ to index time, etc.

We use x to represent an observed data vector. In a supervised problem, we use y or $\mathbf{y}$ to represent the desired output label. We use $\mathbf{z}$ to represent a hidden variable. Sometimes we also use q to represent a hidden discrete variable.

Symbol	Meaning
C	Number of classes
M	Number of data cases
M_c	Number of examples of class c,$M_c = \sum_{i=1}^{M} \mathbb{I}(y_i = c)$
N	Dimensionality of data vector (number of features)
R	Number of outputs (response variables)
$\mathcal{D}$	Training data $\mathcal{D} = \{(\mathbf{x}_i, y_i) \| i = 1 : M\}$
$\mathcal{D}_{test}$	Test data
$\mathcal{X}$	Input space
$\mathcal{Y}$	Output space
K	Number of states or dimensions of a variable (often latent)
$k(x,y)$	Kernel function
$\mathbf{K}$	Kernel matrix
$\mathcal{H}$	Hypothesis space

L	Loss function
$J(\theta)$	Cost function
$f(\mathbf{x})$	Decision function
$P(y \mid \mathbf{x})$	TODO
λ	Strength of ℓ_2 or $\ell_1 regularizer$
$\phi(x)$	Basis function expansion of feature vector $\mathbf{x}$
Φ	Basis function expansion of design matrix $\mathbf{X}$
$q()$	Approximate or proposal distribution
$Q(\theta, \theta_{old})$	Auxiliary function in EM
T	Length of a sequence
$T(\mathcal{D})$	Test statistic for data
$\mathbf{T}$	Transition matrix of Markov chain
θ	Parameter vector
$\theta^{(s)}$	s'th sample of parameter vector
$\hat{\theta}$	Estimate (usually MLE or MAP) of θ
$\hat{\theta}_{MLE}$	Maximum likelihood estimate of θ
$\hat{\theta}_{MAP}$	MAP estimate of θ
$\bar{\theta}$	Estimate (usually posterior mean) of θ
$\mathbf{w}$	Vector of regression weights (called β in statistics)
b	intercept (called ε in statistics)
$\mathbf{W}$	Matrix of regression weights
x_{ij}	Component (i.e., feature) j of data case i ,for $i = 1:N, j = 1:D$
$\mathbf{x}_i$	Training case, $i = 1:N$
$\mathbf{X}$	Design matrix of size $N \times D$
$\bar{\mathbf{x}}$	Empirical mean $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$
$\tilde{\mathbf{x}}$	Future test case
$\mathbf{x}_*$	Future test case
$\mathbf{y}$	Vector of all training labels $\mathbf{y} = (y_1, ..., y_N)$
z_{ij}	Latent component j for case i

Chapter 1
Introduction

1.1 Types of machine learning

$$Machine\,learning\begin{cases}Supervised\,learning\begin{cases}Classfication\\Regression\end{cases}\\Unsupervised\,learning\begin{cases}Discovering\,clusters\\Discovering\,latent\,factors\\Discovering\,graph\,structure\\Matrix\,completion\end{cases}\end{cases}$$

1.2 Three elements of a machine learning method

method = model + strategy + algorithm

1.2.1 Model

In supervised learning, a model is a decision function or conditional probability distribution to be learned. The model's hypothesis space contains all possible decition fuctions $f(x)$ or conditional probability distributions $P(y|\mathbf{x})$.

1.2.2 Strategy

Given a model's hypothesis space, we need a strategy to select which hypothesis is optimal.

1.2.2.1 Loss function and risk function

Definition 1.1. In order to measure how well a function fits the training data, a **loss function** $L : Y \times Y \rightarrow R \geq 0$ is defined. For training example (x_i, y_i), the loss of predicting the value $\widehat{y}$ is $L(y_i, \widehat{y})$.

The following is some common loss functions:

1. 0-1 loss function $L(Y, f(X)) = I(Y, f(X)) = \begin{cases} 1, & Y = f(X) \\ 0, & Y \neq f(X) \end{cases}$
2. Quadratic loss function $L(Y, f(X)) = (Y - f(X))^2$
3. Absolute loss function $L(Y, f(X)) = |Y - f(X)|$
4. Logarithmic loss function $L(Y, P(Y|X)) = -\log P(Y|X)$

Definition 1.2. The risk of function f is defined as the expected loss of f:

$$R_{exp}(f) = E_p\left[L\left(Y, f(X)\right)\right] = \int_{X \times Y} L\left(y, f(x)\right) P(x, y) dx dy \tag{1.1}$$

which is also called expected loss or **risk function**.

Definition 1.3. The risk function $R_{exp}(f)$ can be estimated from the training data as

$$R_{emp}(f) = \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, f(x_i)\right) \tag{1.2}$$

which is also called empirical loss or **empirical risk**.

You can define your own loss function, but if you're a novice, you're probably better off using one from the literature. There are conditions that loss functions should meet[1]:

1. They should approximate the actual loss you're trying to minimize. As was said in the other answer, the standard loss functions for classification is zero-one-loss (misclassification rate) and the ones used for training classifiers are approximations of that loss.
2. The loss function should work with your intended optimization algorithm. That's why zero-one-loss is not used directly: it doesn't work with gradient-based optimization methods since it doesn't have a well-defined gradient (or even a subgradient, like the hinge loss for SVMs has).
   The main algorithm that optimizes the zero-one-loss directly is the old perceptron algorithm(chapter §3).

[1] http://t.cn/zTrDxLO

1.2.2.2 ERM and SRM

Definition 1.4. ERM(Empirical risk minimization)

$$\min_{f\in\mathcal{F}} R_{emp}(f) = \min_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) \tag{1.3}$$

Definition 1.5. Structural risk

$$R_{smp}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \tag{1.4}$$

Definition 1.6. SRM(Structural risk minimization)

$$\min_{f\in\mathcal{F}} R_{srm}(f) = \min_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \tag{1.5}$$

1.2.3 Algorithm

Namely training algorithm(or learning algorithm), which is used to compute the optimal result according to the strategy. It's a procedural concept.

1.3 Cross validation

Definition 1.7. Cross validation, sometimes called *rotation estimation*, is a *model validation* technique for assessing how the results of a statistical analysis will generalize to an independent data set[2].

Common types of cross-validation:

1. K-fold cross-validation. In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k 1 subsamples are used as training data.
2. 2-fold cross-validation. Also, called simple cross-validation or holdout method. This is the simplest variation of k-fold cross-validation, k=2.
3. Leave-one-out cross-validation(*LOOCV*). k=M, the number of original samples.

[2] http://en.wikipedia.org/wiki/Cross-validation_(statistics)

1.4 Linear Regression

Given

$$\begin{aligned}
&\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1 : M\} \\
&\mathcal{H} = \left\{ f(\mathbf{x_i}) = \mathbf{w}^T \mathbf{x_i} + b | i = 1 : M \right\} \\
&L(\mathbf{w}, b) = \sum_{i=1}^{M} (y_i - f(\mathbf{x}_i) - b)^2
\end{aligned} \tag{1.6}$$

Let $\widehat{\mathbf{w}} = \left(\mathbf{w}^T, b\right)^T$, and

$$\widehat{\mathbf{X}} = \begin{pmatrix} \widehat{\mathbf{x}}_1^T \\ \widehat{\mathbf{x}}_2^T \\ \vdots \\ \widehat{\mathbf{x}}_M^T \end{pmatrix}, where\ \widehat{\mathbf{x}}_i = \left(\mathbf{x}_i^T, 1\right)^T \tag{1.7}$$

We can get

$$\begin{aligned}
&L(\widehat{\mathbf{w}}) = \left(\mathbf{y} - \widehat{\mathbf{X}}\widehat{\mathbf{w}}\right)^T \left(\mathbf{y} - \widehat{\mathbf{X}}\widehat{\mathbf{w}}\right) \\
&\frac{\partial L}{\partial \widehat{\mathbf{w}}} = -2\widehat{\mathbf{X}}^T \mathbf{y} + 2\widehat{\mathbf{X}}^T \widehat{\mathbf{X}}\widehat{\mathbf{w}} = 0 \\
&\widehat{\mathbf{X}}^T \mathbf{y} = \widehat{\mathbf{X}}^T \widehat{\mathbf{X}}\widehat{\mathbf{w}} \\
&\widehat{\mathbf{w}} = \left(\widehat{\mathbf{X}}^T \widehat{\mathbf{X}}\right)^{-1} \widehat{\mathbf{X}}^T \mathbf{y}
\end{aligned} \tag{1.8}$$

If $\widehat{\mathbf{X}}^T \widehat{\mathbf{X}}$ is singular, the pseudo-inverse can be used, or else the technique of ridge regression described below can be applied.

Chapter 2
Probability

2.1 Frequentists vs. Bayesians

what is probability? **frequentist** interpretation vs. **Bayesian** interpretation.

One big advantage of the Bayesian interpretation is that it can be used to model our uncertainty about events that do not have long term frequencies.

Therefore most machine learning books adopt the Bayesian interpretation. Fortunately, the basic rules of probability theory are the same, no matter which interpretation is adopted.

2.2 A brief review of probability theory

2.2.1 Basic concepts

We denote a random event by defining a **random variable** X.

Descrete random variable: X can take on any value from a finite or countably infinite set.

Continuous random variable: the value of X is real-valued.

2.2.1.1 CDF

$$F(x) \triangleq P(X \leq x) = \begin{cases} \sum_{u \leq x} p(u), & \text{descrete random variable} \\ \int_{-\infty}^{x} f(u)du, & \text{continuous random variable} \end{cases} \tag{2.1}$$

2.2.1.2 PMF and PDF

For descrete random variable, We denote the probability of the event that $X = x$ by $P(X = x)$, or just $p(x)$ for short. Here $p(x)$ is called a **probability mass function** or **PMF**.A probability mass function is a function that gives the probability that a discrete random variable is exactly equal to some value[3]. This satisfies the properties $0 \leq p(x) \leq 1$ and $\sum_{x \in \mathcal{X}} p(x) = 1$.

For continuous variable, in the equation $F(x) = \int_{-\infty}^{x} f(u)du$, the function $f(x)$ is called a **probability density function** or **PDF**. A probability density function is a function that describes the relative likelihood for this random variable to take on a given value[4].This satisfies the properties $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$.

2.2.2 Mutivariate random variables

2.2.2.1 Joint CDF

We denote joint CDF by $F(x,y) \triangleq P(X \leq x \cap Y \leq y) = P(X \leq x, Y \leq y)$.

$$F(x,y) \triangleq P(X \leq x, Y \leq y) = \begin{cases} \sum_{u \leq x, v \leq y} p(u,v), & \text{descrete} \\ \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v)dudv, & \text{continuous} \end{cases} \tag{2.2}$$

product rule:

$$p(X,Y) = P(X|Y)P(Y) \tag{2.3}$$

Chain rule:

$$p(X_{1:N}) = p(X_1)p(X_2|X_1)p(X_3|X_2,X_1)...p(X_N|X_{1:N-1}) \tag{2.4}$$

2.2.2.2 Marginal distribution

Marginal CDF:

$$\left\{ \begin{array}{l} F_X(x) \triangleq F(x,+\infty) = \begin{cases} \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} \sum_{j=1}^{+\infty} P(X = x_i, Y = y_j), & \text{descrete} \\ \int_{-\infty}^{x} f_X(u)du = \int_{-\infty}^{x} \int_{-\infty}^{+\infty} f(u,v)dudv, & \text{continuous} \end{cases} \\ F_Y(y) \triangleq F(+\infty,y) = \begin{cases} \sum_{y_j \leq y} p(Y = y_j) = \sum_{i=1}^{+\infty} \sum_{y_j \leq y} P(X = x_i, Y = y_j), & \text{descrete} \\ \int_{-\infty}^{y} f_Y(v)dv = \int_{-\infty}^{+\infty} \int_{-\infty}^{y} f(u,v)dudv, & \text{continuous} \end{cases} \end{array} \right. \tag{2.5}$$

Marginal PMF or PDF:

[3] http://en.wikipedia.org/wiki/Probability_mass_function

[4] http://en.wikipedia.org/wiki/Probability_density_function

$$\begin{cases} \begin{cases} P(X=x_i)=\sum_{j=1}^{+\infty}P(X=x_i,Y=y_j), & \text{descrete} \\ f_X(x)=\int_{-\infty}^{+\infty}f(x,y)dy, & \text{continuous} \end{cases} \\ \begin{cases} p(Y=y_j)=\sum_{i=1}^{+\infty}P(X=x_i,Y=y_j), & \text{descrete} \\ f_Y(y)=\int_{-\infty}^{+\infty}f(x,y)dx, & \text{continuous} \end{cases} \end{cases} \tag{2.6}$$

2.2.2.3 Conditional distribution

Conditional PMF:

$$p(X=x_i|Y=y_j)=\frac{p(X=x_i,Y=y_j)}{p(Y=y_j)} \text{ if } p(Y)>0 \tag{2.7}$$

The pmf $p(X|Y)$ is called **conditional probability**.

Conditional PDF:

$$f_{X|Y}(x|y)=\frac{f(x,y)}{f_Y(y)} \tag{2.8}$$

2.2.3 Bayes rule

$$p(Y=y|X=x)=\frac{p(X=x,Y=y)}{p(X=x)}=\frac{p(X=x|Y=y)p(Y=y)}{\sum_{y'}p(X=x|Y=y')p(Y=y')} \tag{2.9}$$

2.2.3.1 Example: Generative classifiers

$$p(y=c|X,\theta)=\frac{p(X|y=c,\theta)p(y=c|\theta)}{\sum_{c'}p(X|y=c',\theta)p(y=c'|\theta)} \tag{2.10}$$

This is called a **generative classifier**, since it specifies how to generate the data using the class conditional density $p(x|y=c)$ and the class prior $p(y=c)$. An alternative approach is to directly fit the class posterior, $p(y=c|x)$;this is known as a **discriminative classifier**.

2.2.4 Independence and conditional independence

We say X and Y are unconditionally independent or marginally independent, denoted $X \perp Y$, if we can represent the joint as the product of the two marginals, i.e.,

$$X \perp Y = P(X,Y)=P(X)P(Y) \tag{2.11}$$

We say X and Y are conditionally independent(CI) given Z if the conditional joint can be written as a product of conditional marginals:

$$X \perp Y|Z = P(X,Y|Z) = P(X|Z)P(Y|Z) \tag{2.12}$$

2.2.5 Quantiles

Since the cdf F is a monotonically increasing function, it has an inverse; let us denote this by F^{-1}. If F is the cdf of X , then $F^{-1}(\alpha)$ is the value of x_α such that $P(X \leq x_\alpha) = \alpha$; this is called the α quantile of F. The value $F^{-1}(0.5)$ is the **median** of the distribution, with half of the probability mass on the left, and half on the right. The values $F^{-1}(0.25)$ and $F^{1}(0.75)$are the lower and upper **quartiles**.

2.2.6 Mean and variance

The most familiar property of a distribution is its **mean**,or **expected value**, denoted by μ. For discrete rvs, it is defined as $\mathbb{E}[X] \triangleq \sum_{x\in\mathcal{X}} xp(x)$, and for continuous rvs, it is defined as $\mathbb{E}[X] \triangleq \int_{\mathcal{X}} xp(x)dx$. If this integral is not finite, the mean is not defined (we will see some examples of this later).

The **variance** is a measure of the spread of a distribution, denoted by σ^2. This is defined as follows:

$$var[X] = \mathbb{E}[(X-\mu)^2] = \int (x-\mu)^2 p(x)dx \tag{2.13}$$

$$= \int x^2 p(x)dx \int \mu^2 p(x)dx - 2\mu \int xp(x)dx = \mathbb{E}[X^2] - \mu^2 \tag{2.14}$$

from which we derive the useful result

$$\mathbb{E}[X^2] = \sigma^2 + \mu^2 \tag{2.15}$$

The **standard deviation** is defined as

$$std[X] \triangleq= \sqrt{var[X]} \tag{2.16}$$

This is useful since it has the same units as X itself.

2.3 Some common discrete distributions

In this section, we review some commonly used parametric distributions defined on discrete state spaces, both finite and countably infinite.

2.3.1 The binomial and Bernoulli distributions

2.3.2 The multinomial and multinoulli distributions

2.3.3 The Poisson distribution

2.3.4 The empirical distribution

2.4 Some common continuous distributions

In this section we present some commonly used univariate (one-dimensional) continuous probability distributions.

2.4.1 Gaussian (normal) distribution

2.4.2 Degenerate pdf

2.4.3 The Laplace distribution

2.4.4 The gamma distribution

2.4.5 The beta distribution

2.4.6 Pareto distribution

2.5 Information theory

2.5.1 Entropy

The entropy of a random variable X with distribution p, denoted by $\mathbb{H}(X)$ or sometimes $\mathbb{H}(p)$, is a measure of its uncertainty. In particular, for a discrete variable with K states, it is defined by

$$\mathbb{H}(X) \triangleq -\sum_{k=1}^{K} p(X=k) \log_2 p(X=k) \tag{2.17}$$

Usually we use log base 2, in which case the units are called **bits**(short for binary digits). If we use log base e , the units are called **nats**.

The discrete distribution with maximum entropy is the uniform distribution (see Section XXX for a proof). Hence for a K-ary random variable, the entropy is maximized if $p(x=k)=1/K$; in this case, $\mathbb{H}(X)=\log_2 K$.

Conversely, the distribution with minimum entropy (which is zero) is any **delta-function** that puts all its mass on one state. Such a distribution has no uncertainty.

2.5.2 KL divergence

One way to measure the dissimilarity of two probability distributions, p and q , is known as the **Kullback-Leibler divergence**(**KL divergence**)or **relative entropy**. This is defined as follows:

$$\mathbb{KL}(P||Q) \triangleq \sum_x p(x)\log_2 \frac{p(x)}{q(x)} \tag{2.18}$$

where the sum gets replaced by an integral for pdfs[5]. The KL divergence is only defined if P and Q both sum to 1 and if $q(x)=0$ implies $p(x)=0$ for all x(absolute continuity). If the quantity $0\ln 0$ appears in the formula, it is interpreted as zero because $\lim\limits_{x\to 0} x\ln x$. We can rewrite this as

$$\mathbb{KL}(p||q) \triangleq \sum_x p(x)\log_2 p(x) - \sum_{k=1}^{K} p(x)\log_2 q(x) = \mathbb{H}(p) - \mathbb{H}(p,q) \tag{2.19}$$

where $\mathbb{H}(p,q)$ is called the **cross entropy**,

$$\mathbb{H}(p,q) = \sum_x p(x)\log_2 q(x) \tag{2.20}$$

One can show (Cover and Thomas 2006) that the cross entropy is the average number of bits needed to encode data coming from a source with distribution p when we use model q to define our codebook. Hence the regular entropy $\mathbb{H}(p)=\mathbb{H}(p,p)$, defined in section §2.5.1,is the expected number of bits if we use the true model, so the KL divergence is the diference between these. In other words, the KL divergence is the average number of *extra* bits needed to encode the data, due to the fact that we used distribution q to encode the data instead of the true distribution p.

The extra number of bits interpretation should make it clear that $\mathbb{KL}(p||q)\geq 0$, and that the KL is only equal to zero if $q=p$. We now give a proof of this important result.

[5] The KL divergence is not a distance, since it is asymmetric. One symmetric version of the KL divergence is the **Jensen-Shannon divergence**, defined as $JS(p_1,p_2)=0.5\mathbb{KL}(p_1||q)+0.5\mathbb{KL}(p_2||q)$,where $q=0.5p_1+0.5p_2$

Theorem 2.1. ***(Information inequality)*** $\mathbb{KL}(p||q) \geq 0$ *with equality iff* $p = q$.

One important consequence of this result is that *the discrete distribution with the maximum entropy is the uniform distribution.*

2.5.3 Mutual information

Definition 2.1. Mutual information or **MI**, is defined as follows:

$$\mathbb{I}(X;Y) \triangleq \mathbb{KL}(P(X,Y)||P(X)P(X)) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2.21}$$

We have $\mathbb{I}(X;Y) \geq 0$ with equality if $P(X,Y) = P(X)P(Y)$. That is, the MI is zero if the variables are independent.

To gain insight into the meaning of MI, it helps to re-express it in terms of joint and conditional entropies. One can show that the above expression is equivalent to the following:

$$\begin{aligned}
\mathbb{I}(X;Y) &= \mathbb{H}(X) - \mathbb{H}(X|Y) && (2.22)\\
&= \mathbb{H}(Y) - \mathbb{H}(Y|X) && (2.23)\\
&= \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) && (2.24)\\
&= \mathbb{H}(X,Y) - \mathbb{H}(X|Y) - \mathbb{H}(Y|X) && (2.25)
\end{aligned}$$

where $\mathbb{H}(X)$ and $\mathbb{H}(Y)$ are the **marginal entropies**, $\mathbb{H}(X|Y)$ and $\mathbb{H}(Y|X)$ are the **conditional entropies**, and $\mathbb{H}(X,Y)$ is the **joint entropy** of X and Y, see Fig. 2.1[6].

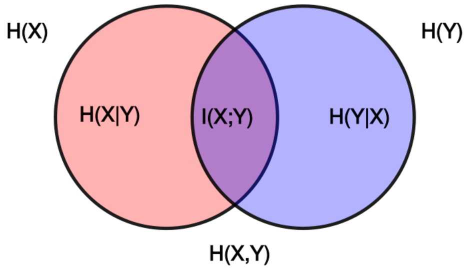

Fig. 2.1 Individual $\mathbb{H}(X), \mathbb{H}(Y)$, joint $\mathbb{H}(X,Y)$, and conditional entropies for a pair of correlated subsystems X, Y with mutual information $\mathbb{I}(X;Y)$.

[6] http://en.wikipedia.org/wiki/Mutual_information

Intuitively, we can interpret the MI between X and Y as the reduction in uncertainty about X after observing Y, or, by symmetry, the reduction in uncertainty about Y after observing X.

A quantity which is closely related to MI is the **pointwise mutual information** or **PMI**. For two events (not random variables) x and y, this is defined as

$$PMI(x,y) \triangleq \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} \tag{2.26}$$

This measures the discrepancy between these events occuring together compared to what would be expected by chance. Clearly the MI of X and Y is just the expected value of the PMI. Interestingly, we can rewrite the PMI as follows:

$$PMI(x,y) = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} \tag{2.27}$$

This is the amount we learn from updating the prior $p(x)$ into the posterior $p(x|y)$, or equivalently, updating the prior $p(y)$ into the posterior $p(y|x)$.

Chapter 3
Perceptron

3.1 Model

$$\mathcal{H}: f(\mathbf{x}) = \text{sign}(\mathbf{w}\mathbf{x} + b) \tag{3.1}$$

where $\text{sign}(x) = \begin{cases} +1, & x \geq 0 \\ -1, & x < 0 \end{cases}$, see Fig. 3.1[7].

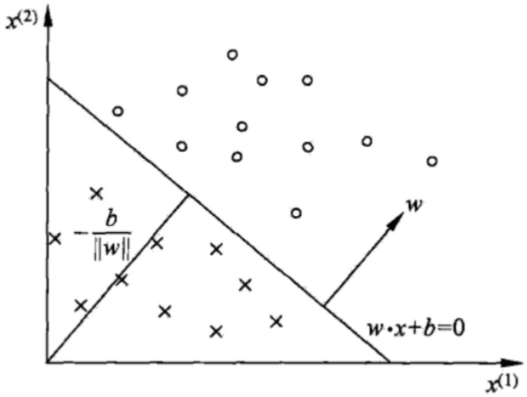

Fig. 3.1 Perceptron

Perceptron is a binary linear classifier, which is a discriminant model.

3.2 Strategy

[7] https://en.wikipedia.org/wiki/Perceptron

$$
\begin{aligned}
L(\mathbf{w},b) &= -y_i(\mathbf{w}\mathbf{x}_i+b) && (3.2)\\
R_{emp}(f) &= -\sum_i y_i(\mathbf{w}\mathbf{x}_i+b) && (3.3)\\
& && (3.4)
\end{aligned}
$$

3.3 Learning algorithm

3.3.1 Primal form

Stochastic gradient descent, the pseudo code is as follows:

```
$\mathbf{w} \leftarrow 0$; $b \leftarrow 0$; $k \leftarrow 0$;
while no mistakes made within the for loop do
    for $i \leftarrow 1$ to $N$ do
        if $y_i(\mathbf{w}^T\mathbf{x}_i + b) \leq 0$ then
            $\mathbf{w} \leftarrow \mathbf{w} + \eta y_i \mathbf{x}_i$;
            $b \leftarrow b + \eta y_i$;
            $k \leftarrow k + 1$;
        end
    end
end
```

Algorithm 1: Perceptron learning algorithm, primal form

3.3.2 Convergency

Theorem 3.1. *(Novikoff) If traning data set $\mathcal{D}$ is linearly separable, then*

1. *There exists a hyperplane denoted as $\widehat{\mathbf{w}}_{opt} \cdot \mathbf{x} + b_{opt} = 0$ which can correctly seperate all samples, and $\exists\gamma > 0, \forall i, y_i(\mathbf{w}_{opt} \cdot \mathbf{x}_i + b_{opt}) \geq \gamma$*
2. *$k \leq \left(\dfrac{R}{\gamma}\right)^2$, where $R = \max\limits_{1\leq i\leq N} ||\widehat{\mathbf{x}}_i||$*

Proof. (1) let $\gamma = \min\limits_i y_i(\mathbf{w}_{opt} \cdot \mathbf{x}_i + b_{opt})$, then we get $y_i(\mathbf{w}_{opt} \cdot \mathbf{x}_i + b_{opt}) \geq \gamma$.

(2) The algorithm start from $\widehat{\mathbf{x}_0} = 0$, if a instance is misclassified, then update the weight. Let $\widehat{\mathbf{w}_{k-1}}$ denotes the extended weight before the k-th misclassified instance, then we can get

$$
\begin{aligned}
y_i(\widehat{\mathbf{w}}_{k-1} \cdot \widehat{\mathbf{x}}_i) &= y_i(\mathbf{w}_{k-1} \cdot \mathbf{x}_i + b_{k-1}) \leq 0 && (3.5)\\
\widehat{\mathbf{w}}_k &= \widehat{\mathbf{w}}_{k-1} + \eta y_i \widehat{\mathbf{x}}_i && (3.6)
\end{aligned}
$$

We could infer the following two equations, the proof procedure are omitted.

1. $\widehat{\mathbf{w}}_k \cdot \widehat{\mathbf{w}}_{opt} \geq k\eta\gamma$
2. $||\widehat{\mathbf{w}}_k||^2 \leq k\eta^2 R^2$

From above two equations we get

$$\begin{aligned}
&k\eta\gamma \leq \widehat{\mathbf{w}}_k \cdot \widehat{\mathbf{w}}_{opt} \leq ||\widehat{\mathbf{w}}_k|| \left|\left|\widehat{\mathbf{w}}_{opt}\right|\right| \leq \sqrt{k}\eta R \\
&k^2\gamma^2 \leq kR^2 \\
&\text{i.e. } k \leq \left(\frac{R}{\gamma}\right)^2
\end{aligned}$$

3.3.3 Dual form

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \tag{3.7}$$

$$b = \sum_{i=1}^{N} \alpha_i y_i \tag{3.8}$$

$$f(\mathbf{x}) = \text{sign}\left(\sum_{j=1}^{N} \alpha_j y_j \mathbf{x}_j \cdot \mathbf{x} + b\right) \tag{3.9}$$

```
$\alpha \leftarrow 0$; $b \leftarrow 0$; $k \leftarrow 0$;
while no mistakes made within the for loop do
    for $i \leftarrow 1$ to $N$ do
        if $y_i \left(\sum_{j=1}^{N} \alpha_j y_j \mathbf{x}_j \cdot \mathbf{x}_i + b\right) \leq 0$ then
            $\alpha \leftarrow \alpha + \eta$;
            $b \leftarrow b + \eta y_i$;
            $k \leftarrow k + 1$;
        end
    end
end
```

Algorithm 2: Perceptron learning algorithm, dual form

Glossary

feture vector A feture vector to represent one data.

loss function a function that maps an event onto a real number intuitively representing some ”cost” associated with the event.

glossary term Write here the description of the glossary term. Write here the description of the glossary term. Write here the description of the glossary term.