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

المرجع السريع في علم تعلّم الآلة

2 جويلية 2017

المحتويات

List of Contributors

Wei Zhang
PhD candidate at the Institute of Software, Chinese Academy of Sciences (ISCAS), Beijing, P.R.CHINA, e-mail: zh3feng@gmail.com, has written chapters of Naive Bayes and SVM.

Fei Pan
Master at Beijing University of Technology, Beijing, P.R.CHINA, e-mail: example@gmail.com, has written chapters of KMeans, AdaBoost.

Yong Li
PhD candidate at the Institute of Automation of the Chinese Academy of Sciences (CASIA), Beijing, P.R.CHINA, e-mail: liyong3forever@gmail.com, has written chapters of Logistic Regression.

Jiankou Li
PhD candidate at the Institute of Software, Chinese Academy of Sciences (ISCAS), Beijing, P.R.CHINA, e-mail: lijiankoucoco@163.com, has written chapters of BayesNet.

Notation

مجموع الرموز

Introduction

تمهيد

من الصعوبة التّوصل إلي مجموعة وحيدة و ثابتة من الرموز لتغطية المجال الشاسع من البيانات (data) و النماذج (models) و الخوارزميات (algorithms) التي نناقشها في هذا الكتيّب. علاوة على ذلك، العلامات الرياضية المتفق عليها تختلف بين علم تعلّم الآلة (machine learning) و علم الإحصاءات (statistics)، و بين الكتب والأوراق العلميّة المختلفة. مع ذلك، فقد حاولنا أن تكون الرموز المستعملة متّسقة قدر الإمكان. فيما يلي نلخّص معظم الرموز المستخدمة, هذا لا ينفي أن بعض المقاطع الفرديّة في الكتيّب قد تعرض رموزا جديدة. إعلم أيضا أن بعض الرموز قد يكون لها معان مختلفة تبعا للسياق, رغم أننا سنحرص على تجنّب ذلك قدر الإمكان.

General math notation

مجموع الرموز الرياضية

Symbol	Meaning
$\lfloor x \rfloor$	Floor of x, i.e., round down to nearest integer
$\lceil x \rceil$	Ceiling of x, i.e., round up to nearest integer
$\boldsymbol{x} \otimes \boldsymbol{y}$	Convolution of $\boldsymbol{x}$ and $\boldsymbol{y}$
$\boldsymbol{x} \odot \boldsymbol{y}$	Hadamard (elementwise) product of $\boldsymbol{x}$ and $\boldsymbol{y}$
$a \wedge b$	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
$\arg\max\limits_x f(x)$	Argmax: the value x that maximizes f
$B(a, b)$	Beta function, $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$
$B(\boldsymbol{\alpha})$	Multivariate beta function, $\frac{\prod\limits_k \Gamma(\alpha_k)}{\Gamma(\sum\limits_k \alpha_k)}$
$\binom{n}{k}$	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, $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} \mathrm{d}u$
$\Psi(x)$	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 رموز علم الجبر الخطي

خلال هذا الكتاب, سنستخدم حروف boldface الصغيرة للدلالة على النّاقلات(vectors) مثل $\boldsymbol{x}$ و حروف boldface العلوية لدلالة على المصفوفات مثل $\boldsymbol{X}$. نشير إلى إدخلات مصفوفة (matrix entries) بأحرف كبيرة غير جريئة، مثل X_{ij}.

سنعتبر كل النّاقلات (vectors) ناقلات عمود (column vectors), مالم يذكر خلاف ذلك في السياق. نستخدم $(x_1, \cdots, x_D)$ للدّلالة على متجه عمود تم إنشاؤه بواسطة تكديس D أعداد (stacking D scalars). إذا كتبنا $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$, حيث الجانب الأيسر هو مصفوفة(matrix) , فإننا نقصد تكديس $\boldsymbol{x}_i$ على طول الأعمدة لخلق مصفوفة(matrix).

Symbol	Meaning
$\boldsymbol{X} \succ 0$	$\boldsymbol{X}$ is a positive definite matrix
$tr(\boldsymbol{X})$	Trace of a matrix
$det(\boldsymbol{X})$	Determinant of matrix $\boldsymbol{X}$
$\lvert\boldsymbol{X}\rvert$	Determinant of matrix $\boldsymbol{X}$
$\boldsymbol{X}^{-1}$	Inverse of a matrix
$\boldsymbol{X}^{\dagger}$	Pseudo-inverse of a matrix
$\boldsymbol{X}^T$	Transpose of a matrix
$\boldsymbol{x}^T$	Transpose of a vector
$\mathrm{diag}(x)$	Diagonal matrix made from vector $\boldsymbol{x}$
$\mathrm{diag}(X)$	Diagonal vector extracted from matrix $\boldsymbol{X}$
$\boldsymbol{I}$ or $\boldsymbol{I}_d$	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)
$\lvert\lvert\boldsymbol{x}\rvert\rvert = \lvert\lvert\boldsymbol{x}\rvert\rvert_2$	Euclidean or ℓ_2 norm $\sqrt{\sum_{j=1}^{d} x_j^2}$
$\lvert\lvert\boldsymbol{x}\rvert\rvert_1$	ℓ_1 norm $\sum_{j=1}^{d} \lvert x_j \rvert$
$\boldsymbol{X}_{:,j}$	j'th column of matrix
$\boldsymbol{X}_{i,:}$	transpose of i'th row of matrix (a column vector)
$\boldsymbol{X}_{i,j}$	Element (i,j) of matrix $\boldsymbol{X}$
$\boldsymbol{x} \otimes \boldsymbol{y}$	Tensor product of $\boldsymbol{x}$ and $\boldsymbol{y}$

Probability notation رموز علم الإحتمال

نرمز إلى الأعداد العشوائية و الثابتة (random and fixed scalars) بخط صغير (lower case), و الناقلات العشوائية و الثابتة (random and fixed vectors) بالحروف الصغيرة الجريئة(bold lower case) و المصفوفات العشوائية و الثابتة (random and fixed matrices) بالحروف الجريئة العلوية (bold upper case). أحيانا نستخدم الحروف العلوية غير الجريئة (non-bold upper case) لدلالة على المتغيرات العددية العشوائية (scalar random variables). نستخدم, أيضا, $p()$ لكل من المتغيرات العشوائية المنفصلة و المستمرة (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
$\mathrm{cov}[X]$	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 $\Lambda = \Sigma^{-1}$
$\mathrm{mode}[\boldsymbol{X}]$	Most probable value of $\boldsymbol{X}$
μ	Mean of a scalar distribution
$\boldsymbol{\mu}$	Mean of a multivariate distribution
Φ	cdf of standard normal
ϕ	pdf of standard normal
$\boldsymbol{\pi}$	multinomial parameter vector, Stationary distribution of Markov chain
ρ	Correlation coefficient
$\mathrm{sigm}(x)$	Sigmoid (logistic) function, $\frac{1}{1+e^{-x}}$
σ^2	Variance
Σ	Covariance matrix
$\mathrm{var}[x]$	Variance of x
ν	Degrees of freedom parameter
Z	Normalization constant of a probability distribution

Machine learning/statistics notation رموز علم تعلّم الآلة و الإحصاءات

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 $\boldsymbol{y}$ to represent the desired output label. We use $\boldsymbol{z}$ to represent a hidden variable. Sometimes we also use q to represent a hidden discrete variable.

Symbol	Meaning
C	Number of classes
D	Dimensionality of data vector (number of features)
N	Number of data cases
N_c	Number of examples of class c,$N_c = \sum_{i=1}^{N} \mathbb{I}(y_i = c)$
R	Number of outputs (response variables)
$\mathcal{D}$	Training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \| i = 1:N\}$
$\mathcal{D}_{test}$	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
$\boldsymbol{K}$	Kernel matrix
$\mathcal{H}$	Hypothesis space
L	Loss function
$J(\boldsymbol{\theta})$	Cost function
$f(\boldsymbol{x})$	Decision function
$P(y\|\boldsymbol{x})$	Conditional probability
λ	Strength of ℓ_2 or $\ell_1 regularizer$
$\phi(x)$	Basis function expansion of feature vector $\boldsymbol{x}$
$\boldsymbol{\Phi}$	Basis function expansion of design matrix $\boldsymbol{X}$
$q()$	Approximate or proposal distribution
$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{old})$	Auxiliary function in EM
T	Length of a sequence
$T(\mathcal{D})$	Test statistic for data
$\boldsymbol{T}$	Transition matrix of Markov chain
$\boldsymbol{\theta}$	Parameter vector
$\boldsymbol{\theta}^{(s)}$	s'th sample of parameter vector
$\hat{\boldsymbol{\theta}}$	Estimate (usually MLE or MAP) of $\boldsymbol{\theta}$
$\hat{\boldsymbol{\theta}}_{MLE}$	Maximum likelihood estimate of $\boldsymbol{\theta}$
$\hat{\boldsymbol{\theta}}_{MAP}$	MAP estimate of $\boldsymbol{\theta}$
$\bar{\boldsymbol{\theta}}$	Estimate (usually posterior mean) of $\boldsymbol{\theta}$
$\boldsymbol{w}$	Vector of regression weights (called $\boldsymbol{\beta}$ in statistics)
b	intercept (called ε in statistics)
$\boldsymbol{W}$	Matrix of regression weights
x_{ij}	Component (i.e., feature) j of data case i ,for $i = 1:N, j = 1:D$
$\boldsymbol{x}_i$	Training case, $i = 1:N$
$\boldsymbol{X}$	Design matrix of size $N \times D$
$\bar{\boldsymbol{x}}$	Empirical mean $\bar{\boldsymbol{x}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i$
$\tilde{\boldsymbol{x}}$	Future test case
$\boldsymbol{x}_*$	Feature test case
$\boldsymbol{y}$	Vector of all training labels $\boldsymbol{y} = (y_1, ..., y_N)$
z_{ij}	Latent component j for case i

باب 1

Introduction تمهيد

Types of machine learning أنواع تعلّم الآلة

- Supervised learning
  - Classification
  - Regression
- Unsupervised learning
  - Discovering clusters
  - Discovering latent factors
  - Discovering graph structure
  - Matrix completion

المكونات الثلاثة لنماذج تعلّم الآلة

النوذج (Model) = التمثيل (Representation) + التقييم (Evaluation) + التحسين (Optimization) [1]
في ما يلي سيقع تفسير كل مكون من هذه المكونات الثلاثة على حدا.

Representation التمثيل

In supervised learning, a model must be represented as a conditional probability distribution $P(y|\boldsymbol{x})$(usually we call it classifier) or a decision function $f(x)$. The set of classifiers(or decision functions) is called the hypothesis space of the model. Choosing a representation for a model is tantamount to choosing the hypothesis space that it can possibly learn.

Evaluation التقييم

In the hypothesis space, an evaluation function (also called objective function or risk function) is needed to distinguish good classifiers(or decision functions) from bad ones.

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
$$L(Y, f(X)) = \mathbb{I}(Y, f(X)) = \begin{cases} 1, & Y = f(X) \\ 0, & Y \neq f(X) \end{cases}$$
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:

$$R_{\text{exp}}(f) = E\left[L\left(Y, f(X)\right)\right] = \int L\left(y, f(x)\right) P(x, y) \mathrm{d}x\mathrm{d}y \quad (1.1)$$

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

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

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

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[2]:

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 §??).

[1] Model = Representation + Evaluation + Optimization. Domingos, P. A few useful things to know about machine learning. Commun. ACM. 87–78:(10)55 .(2012)

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

Evaluation

No training is needed.

Optimization

No training is needed.

Overfitting

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[3].

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.

Model selection

When we have a variety of models of different complexity (e.g., linear or logistic regression models with different degree polynomials, or KNN classifiers with different values ofK), how should we pick the right one? A natural approach is to compute the misclassification rate on the training set for each method.

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

ERM and SRM

Definition 1.4. ERM(Empirical risk minimization)

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

Definition 1.5. Structural risk

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

Definition 1.6. SRM(Structural risk minimization)

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

Optimization التقييم

Finally, we need a training algorithm(also called learning algorithm) to search among the classifiers in the the hypothesis space for the highest-scoring one. The choice of optimization technique is key to the efficiency of the model.

Some basic concepts

Parametric vs non-parametric models

A simple non-parametric classfiier: K-nearest neighbours

Representation

$$y = f(\boldsymbol{x}) = \arg\min_{c} \sum_{\boldsymbol{x}_i \in N_k(\boldsymbol{x})} \mathbb{I}(y_i = c) \quad (1.6)$$

where $N_k(\boldsymbol{x})$ is the set of k points that are closest to point $\boldsymbol{x}$.

Usually use k-d tree to accelerate the process of finding k nearest points.