

<https://github.com/soulmachine/machine-learning-cheat-sheet>

soulmachine@gmail.com

# Machine Learning Cheat Sheet

Classical equations, diagrams and tricks in machine learning

December 1, 2022

©2013 soulmachine

Except where otherwise noted, This document is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA3.0) license

(<http://creativecommons.org/licenses/by/3.0/>).

# Preface

This cheat sheet is a condensed version of machine learning manual, which contains many classical equations and diagrams on machine learning, and aims to help you quickly recall knowledge and ideas in machine learning.

This cheat sheet has two significant advantages:

1. Clearer symbols. Mathematical formulas use quite a lot of confusing symbols. For example,  $X$  can be a set, a random variable, or a matrix. This is very confusing and makes it very difficult for readers to understand the meaning of math formulas. This cheat sheet tries to standardize the usage of symbols, and all symbols are clearly pre-defined, see section §.
2. Less thinking jumps. In many machine learning books, authors omit some intermediary steps of a mathematical proof process, which may save some space but causes difficulty for readers to understand this formula and readers get lost in the middle way of the derivation process. This cheat sheet tries to keep important intermediary steps as where as possible.



# Contents

<b>Contents</b> .....	v	2.7	Monte Carlo approximation .....	12
<b>Notation</b> .....	xi	2.8	Information theory .....	12
		2.8.1	Entropy .....	12
		2.8.2	KL divergence .....	12
		2.8.3	Mutual information .....	13
<b>1 Introduction</b> .....	1	<b>3 Generative models for discrete data</b> .....	15	
1.1 Types of machine learning .....	1	3.1 Generative classifier .....	15	
1.2 Three elements of a machine learning model .....	1	3.2 Bayesian concept learning .....	15	
1.2.1 Representation .....	1	3.2.1 Likelihood .....	15	
1.2.2 Evaluation .....	1	3.2.2 Prior .....	15	
1.2.3 Optimization .....	2	3.2.3 Posterior .....	15	
1.3 Some basic concepts .....	2	3.2.4 Posterior predictive distribution .	15	
1.3.1 Parametric vs non-parametric models .....	2	3.3 The beta-binomial model .....	16	
1.3.2 A simple non-parametric classifier: K-nearest neighbours .	2	3.3.1 Likelihood .....	16	
1.3.3 Overfitting .....	2	3.3.2 Prior .....	16	
1.3.4 Cross validation .....	2	3.3.3 Posterior .....	16	
1.3.5 Model selection .....	2	3.3.4 Posterior predictive distribution .	16	
<b>2 Probability</b> .....	3	3.4 The Dirichlet-multinomial model .....	17	
2.1 Frequentists vs. Bayesians .....	3	3.4.1 Likelihood .....	17	
2.2 A brief review of probability theory .....	3	3.4.2 Prior .....	17	
2.2.1 Basic concepts .....	3	3.4.3 Posterior .....	17	
2.2.2 Mutivariate random variables .....	3	3.4.4 Posterior predictive distribution .	18	
2.2.3 Bayes rule .....	4	3.5 Naive Bayes classifiers .....	18	
2.2.4 Independence and conditional independence .....	4	3.5.1 Optimization .....	18	
2.2.5 Quantiles .....	4	3.5.2 Using the model for prediction .	19	
2.2.6 Mean and variance .....	4	3.5.3 The log-sum-exp trick .....	19	
2.3 Some common discrete distributions .....	4	3.5.4 Feature selection using mutual information .....	19	
2.3.1 The Bernoulli and binomial distributions .....	5	3.5.5 Classifying documents using bag of words .....	19	
2.3.2 The multinoulli and multinomial distributions .....	5	<b>4 Gaussian Models</b> .....	21	
2.3.3 The Poisson distribution .....	5	4.1 Basics .....	21	
2.3.4 The empirical distribution .....	5	4.1.1 MLE for a MVN .....	21	
2.4 Some common continuous distributions ..	5	4.1.2 Maximum entropy derivation of the Gaussian *	21	
2.4.1 Gaussian (normal) distribution ..	5	4.2 Gaussian discriminant analysis .....	22	
2.4.2 Student's t-distribution .....	6	4.2.1 Quadratic discriminant analysis (QDA) .....	22	
2.4.3 The Laplace distribution .....	6	4.2.2 Linear discriminant analysis (LDA) .....	22	
2.4.4 The gamma distribution .....	7	4.2.3 Two-class LDA .....	23	
2.4.5 The beta distribution .....	7	4.2.4 MLE for discriminant analysis ..	24	
2.4.6 Pareto distribution .....	7	4.2.5 Strategies for preventing overfitting .....	24	
2.5 Joint probability distributions .....	8	4.2.6 Regularized LDA *	24	
2.5.1 Covariance and correlation .....	8	4.2.7 Diagonal LDA .....	24	
2.5.2 Multivariate Gaussian distribution	9	4.2.8 Nearest shrunken centroids classifier *	24	
2.5.3 Multivariate Student's t-distribution .....	9	4.3 Inference in jointly Gaussian distributions	24	
2.5.4 Dirichlet distribution .....	9	4.3.1 Statement of the result .....	25	
2.6 Transformations of random variables .....	10	4.3.2 Examples .....	25	
2.6.1 Linear transformations .....	10	4.4 Linear Gaussian systems .....	25	
2.6.2 General transformations .....	10			
2.6.3 Central limit theorem .....	10			

4.4.1	Statement of the result	25	7.4.3	Connection with PCA *	39
4.5	Digression: The Wishart distribution *	25	7.4.4	Regularization effects of big data	39
4.6	Inferring the parameters of an MVN	25	7.5	Bayesian linear regression	39
4.6.1	Posterior distribution of $\mu$	25	8	<b>Logistic Regression</b>	41
4.6.2	Posterior distribution of $\Sigma$ *	25	8.1	Representation	41
4.6.3	Posterior distribution of $\mu$ and $\Sigma$ *	25	8.2	Optimization	41
4.6.4	Sensor fusion with unknown precisions *	25	8.2.1	MLE	41
5	<b>Bayesian statistics</b>	27	8.2.2	MAP	41
5.1	Introduction	27	8.3	Multinomial logistic regression	41
5.2	Summarizing posterior distributions	27	8.3.1	Representation	41
5.2.1	MAP estimation	27	8.3.2	MLE	42
5.2.2	Credible intervals	28	8.3.3	MAP	42
5.2.3	Inference for a difference in proportions	28	8.4	Bayesian logistic regression	42
5.3	Bayesian model selection	29	8.4.1	Laplace approximation	42
5.3.1	Bayesian Occam's razor	29	8.4.2	Derivation of the BIC	42
5.3.2	Computing the marginal likelihood (evidence)	30	8.4.3	Gaussian approximation for logistic regression	42
5.3.3	Bayes factors	31	8.4.4	Approximating the posterior predictive	42
5.4	Priors	31	8.4.5	Residual analysis (outlier detection) *	42
5.4.1	Uninformative priors	31	8.5	Online learning and stochastic optimization	42
5.4.2	Robust priors	31	8.5.1	The perceptron algorithm	42
5.4.3	Mixtures of conjugate priors	31	8.6	Generative vs discriminative classifiers	44
5.5	Hierarchical Bayes	32	8.6.1	Pros and cons of each approach	44
5.6	Empirical Bayes	32	8.6.2	Dealing with missing data	44
5.7	Bayesian decision theory	32	8.6.3	Fisher's linear discriminant analysis (FLDA) *	45
5.7.1	Bayes estimators for common loss functions	32	9	<b>Generalized linear models and the exponential family</b>	47
5.7.2	The false positive vs false negative tradeoff	33	9.1	The exponential family	47
6	<b>Frequentist statistics</b>	35	9.1.1	Definition	47
6.1	Sampling distribution of an estimator	35	9.1.2	Examples	47
6.1.1	Bootstrap	35	9.1.3	Log partition function	48
6.1.2	Large sample theory for the MLE *	35	9.1.4	MLE for the exponential family	48
6.2	Frequentist decision theory	35	9.1.5	Bayes for the exponential family	49
6.3	Desirable properties of estimators	35	9.1.6	Maximum entropy derivation of the exponential family *	49
6.4	Empirical risk minimization	35	9.2	Generalized linear models (GLMs)	49
6.4.1	Regularized risk minimization	35	9.2.1	Basics	49
6.4.2	Structural risk minimization	35	9.3	Probit regression	49
6.4.3	Estimating the risk using cross validation	35	9.4	Multi-task learning	49
6.4.4	Upper bounding the risk using statistical learning theory *	35	10	<b>Directed graphical models (Bayes nets)</b>	51
6.4.5	Surrogate loss functions	35	10.1	Introduction	51
6.5	Pathologies of frequentist statistics *	35	10.1.1	Chain rule	51
7	<b>Linear Regression</b>	37	10.1.2	Conditional independence	51
7.1	Introduction	37	10.1.3	Graphical models	51
7.2	Representation	37	10.1.4	Directed graphical model	51
7.3	MLE	37	10.2	Examples	51
7.3.1	OLS	37	10.2.1	Naive Bayes classifiers	51
7.3.2	SGD	38	10.2.2	Markov and hidden Markov models	52
7.4	Ridge regression(MAP)	38	10.3	Inference	52
7.4.1	Basic idea	38	10.4	Learning	52
7.4.2	Numerically stable computation *	39	10.4.1	Learning from complete data	52
			10.4.2	Learning with missing and/or latent variables	52

10.5	Conditional independence properties of DGMs	52	12.2.1	Classical PCA	65
10.5.1	d-separation and the Bayes Ball algorithm (global Markov properties)	52	12.2.2	Singular value decomposition (SVD)	66
10.5.2	Other Markov properties of DGMs	53	12.2.3	Probabilistic PCA	67
10.5.3	Markov blanket and full conditionals	53	12.2.4	EM algorithm for PCA	67
10.5.4	Multinoulli Learning	53	12.3	Choosing the number of latent dimensions	68
10.6	Influence (decision) diagrams *	53	12.3.1	Model selection for FA/PPCA	68
			12.3.2	Model selection for PCA	68
11	Mixture models and the EM algorithm	55	12.4	PCA for categorical data	68
11.1	Latent variable models	55	12.5	PCA for paired and multi-view data	68
11.2	Mixture models	55	12.5.1	Supervised PCA (latent factor regression)	68
11.2.1	Mixtures of Gaussians	55	12.5.2	Discriminative supervised PCA	68
11.2.2	Mixtures of multinoullis	55	12.5.3	Canonical correlation analysis	68
11.2.3	Using mixture models for clustering	56	12.6	Independent Component Analysis (ICA)	68
11.2.4	Mixtures of experts	56	12.6.1	Maximum likelihood estimation	69
11.3	Parameter estimation for mixture models	57	12.6.2	The FastICA algorithm	69
11.3.1	Unidentifiability	57	12.6.3	Using EM	69
11.3.2	Computing a MAP estimate is non-convex	57	12.6.4	Other estimation principles *	69
11.4	The EM algorithm	57	13	Sparse linear models	71
11.4.1	Introduction	57	14	Kernels	73
11.4.2	Basic idea	57	14.1	Introduction	73
11.4.3	EM for GMMs	58	14.2	Kernel functions	73
11.4.4	EM for K-means	59	14.2.1	RBF kernels	73
11.4.5	EM for mixture of experts	60	14.2.2	TF-IDF kernels	73
11.4.6	EM for DGMs with hidden variables	60	14.2.3	Mercer (positive definite) kernels	73
11.4.7	EM for the Student distribution *	60	14.2.4	Linear kernels	74
11.4.8	EM for probit regression *	60	14.2.5	Matern kernels	74
11.4.9	Derivation of the $Q$ function	60	14.2.6	String kernels	74
11.4.10	Convergence of the EM Algorithm *	60	14.2.7	Pyramid match kernels	74
11.4.11	Generalization of EM Algorithm *	61	14.2.8	Kernels derived from probabilistic generative models	74
11.4.12	Online EM	62	14.3	Using kernels inside GLMs	75
11.4.13	Other EM variants *	62	14.3.1	Kernel machines	75
11.5	Model selection for latent variable models	62	14.3.2	L1VMs, RVMs, and other sparse vector machines	76
11.5.1	Model selection for probabilistic models	62	14.4	The kernel trick	76
11.5.2	Model selection for non-probabilistic methods	62	14.4.1	Kernelized KNN	76
11.6	Fitting models with missing data	62	14.4.2	Kernelized K-medoids clustering	76
11.6.1	EM for the MLE of an MVN with missing data	62	14.4.3	Kernelized ridge regression	76
			14.4.4	Kernel PCA	76
12	Latent linear models	63	14.5	Support vector machines (SVMs)	77
12.1	Factor analysis	63	14.5.1	SVMs for classification	77
12.1.1	FA is a low rank parameterization of an MVN	63	14.5.2	SVMs for regression	78
12.1.2	Inference of the latent factors	63	14.5.3	Choosing $C$	78
12.1.3	Unidentifiability	63	14.5.4	A probabilistic interpretation of SVMs	78
12.1.4	Mixtures of factor analysers	64	14.5.5	Summary of key points	79
12.1.5	EM for factor analysis models	64	14.6	Comparison of discriminative kernel methods	79
12.1.6	Fitting FA models with missing data	65	14.7	Kernels for building generative models	79
12.2	Principal components analysis (PCA)	65	15	Gaussian processes	81
			15.1	Introduction	81
			15.2	GPs for regression	81
			15.3	GPs meet GLMs	81
			15.4	Connection with other methods	81
			15.5	GP latent variable model	81
			15.6	Approximation methods for large datasets	81

<b>16</b>	<b>Adaptive basis function models</b>	83	<b>24.5</b>	<b>Auxiliary variable MCMC *</b>	99
16.1	AdaBoost	83	<b>25</b>	<b>Clustering</b>	101
16.1.1	Representation	83	<b>26</b>	<b>Graphical model structure learning</b>	103
16.1.2	Evaluation	83	<b>27</b>	<b>Latent variable models for discrete data</b>	105
16.1.3	Optimization	83	27.1	Introduction	105
16.1.4	The upper bound of the training error of AdaBoost	83	27.2	Distributed state LVMs for discrete data	105
<b>17</b>	<b>Hidden markov Model</b>	85	<b>28</b>	<b>Deep learning</b>	107
17.1	Introduction	85	<b>A</b>	<b>Optimization methods</b>	109
17.2	Markov models	85	A.1	Convexity	109
<b>18</b>	<b>State space models</b>	87	A.2	Gradient descent	109
<b>19</b>	<b>Undirected graphical models (Markov random fields)</b>	89	A.2.1	Stochastic gradient descent	109
<b>20</b>	<b>Exact inference for graphical models</b>	91	A.2.2	Batch gradient descent	109
<b>21</b>	<b>Variational inference</b>	93	A.2.3	Line search	109
<b>22</b>	<b>More variational inference</b>	95	A.2.4	Momentum term	109
<b>23</b>	<b>Monte Carlo inference</b>	97	A.3	Lagrange duality	109
<b>24</b>	<b>Markov chain Monte Carlo (MCMC)inference</b>	99	A.3.1	Primal form	109
24.1	Introduction	99	A.3.2	Dual form	110
24.2	Metropolis Hastings algorithm	99	A.4	Newton's method	110
24.3	Gibbs sampling	99	A.5	Quasi-Newton method	110
24.4	Speed and accuracy of MCMC	99	A.5.1	DFP	110
			A.5.2	BFGS	110
			A.5.3	Broyden	110
			<b>Glossary</b>		111



# 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](mailto: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](mailto: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](mailto: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](mailto:lijiankoucoco@163.com), has written chapters of BayesNet.

