

Tue Herlau, Mikkel N. Schmidt and Morten Mørup

# Introduction to Machine Learning and Data Mining

Material for continuing education course, Spring 2019

This document may not be redistributed. All rights belongs to  
the authors and DTU.

February 18, 2019

Technical University of Denmark



---

## Notation cheat sheet

	Matlab var.	Type	Size	Description
	<b>X</b>	Numeric	$N \times M$	Data matrix: The rows correspond to $N$ data objects, each of which contains $M$ attributes.
	<b>attributeNames</b>	Cell array	$M \times 1$	Attribute names: Name (string) for each of the $M$ attributes.
	<b>N</b>	Numeric	Scalar	Number of data objects.
	<b>M</b>	Numeric	Scalar	Number of attributes.
Regression	<b>y</b>	Numeric	$N \times 1$	Dependent variable (output): For each data object, <b>y</b> contains an output value that we wish to predict.
Classification	<b>y</b>	Numeric	$N \times 1$	Class index: For each data object, <b>y</b> contains a class index, $y_n \in \{0, 1, \dots, C - 1\}$ , where $C$ is the total number of classes.
	<b>classNames</b>	Cell array	$C \times 1$	Class names: Name (string) for each of the $C$ classes.
	<b>C</b>	Numeric	Scalar	Number of classes.
Cross-validation				All variables mentioned above appended with <b>_train</b> or <b>_test</b> represent the corresponding variable for the training or test set.
	<b>*_train</b>	—	—	Training data.
	<b>*_test</b>	—	—	Test data.

This book attempts to give a concise introduction to machine-learning concepts. We believe this is best accomplished by clearly stating what a given method actually does as a sequence of mathematical operations, and use illustrations and text to provide an intuition. We will therefore make use of tools from linear algebra, probability theory and analysis to describe the methods, focusing on using as small a set of concepts as possible and strive towards maximal consistency.

## VI

In the following, vectors will be denoted by lower-case roman letters  $\mathbf{x}, \mathbf{y}, \dots$  and matrices by bolder, upper case roman letters  $\mathbf{A}, \mathbf{B}, \dots$ . A superscript  $T$  denote the transpose. For instance

$$\mathbf{A} = \begin{bmatrix} -1 & 0 & 2 \\ 1 & 1 & -2 \end{bmatrix} \text{ and if } \mathbf{x} = \begin{bmatrix} -1 \\ 4 \\ 1 \end{bmatrix} \text{ then } \mathbf{x}^T = [-1 \ 4 \ 1].$$

The  $i$ th element of a vector is written as  $x_i$  and the  $i, j$ 'th element of a matrix as  $A_{ij}$  (and sometimes  $A_{i,j}$  to avoid ambiguity). In the preceding example,  $x_2 = 4$  and  $A_{2,3} = -2$ . During this course the observed data set, which we feed into our machine learning methods, will consist of  $N$  observations where each observation consist of a  $M$  dimensional vector. For instance if we have  $N$  observations  $\mathbf{x}_1, \dots, \mathbf{x}_N$  then any given observation will consist of  $M$  numbers:

$$\mathbf{x} = [x_1 \ \dots \ x_M]^T.$$

For convenience, we will often combine the observations into an  $N \times M$  data matrix  $\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix}$$

in which the  $i$ th row of  $\mathbf{X}$  corresponds to the row vector  $\mathbf{x}_i^T$ . We will use this notation for our data matrix and the *rows* of  $\mathbf{X}$  will correspond to  $N$  *observations* and the  $M$  *columns* of  $\mathbf{X}$  will correspond to  $M$  *attributes*. Often each of the observations  $\mathbf{x}_i$  will come with a *label* or *target*  $y_i$  corresponding to a feature of  $\mathbf{x}_i$  which we are interested in predicting. In this case we will collect the labels in a  $N$ -dimensional vector  $\mathbf{y}$  and the pair  $(\mathbf{X}, \mathbf{y})$  will be all the data available for the machine learning method. A more comprehensive translation of the notation as used in this book and in the exercises can be found in the table on the previous page. Finally, the reader should be familiar with the big-sigma notation which allows us to conveniently write sums and products of multiple terms:

$$\sum_{i=1}^n f(i) = f(1) + f(2) + \dots + f(n-1) + f(n)$$

$$\prod_{i=1}^n f(i) = f(1) \times f(2) \times \dots \times f(n-1) \times f(n).$$

As an example, if  $f(i) = i^2$  and  $n = 4$  we have

$$\sum_{i=1}^4 f(i) = 1^2 + 2^2 + 3^2 + 4^2 = 30, \quad \prod_{i=1}^4 f(i) = 1^2 \times 2^2 \times 3^2 \times 4^2 = 576.$$

---

# Contents

Notation cheat sheet .....	V
----------------------------	---

---

## Part I Data: Types, Features and Visualization

---

<b>1 Introduction</b> .....	3
1.1 What is machine learning and data mining .....	3
1.1.1 Machine Learning .....	3
1.1.2 Data mining .....	4
1.1.3 Relationship to artificial intelligence .....	4
1.1.4 Relationship to other disciplines .....	4
1.1.5 Why should I care about machine learning .....	4
1.2 Machine learning tasks .....	5
1.2.1 Supervised learning .....	5
1.2.2 Unsupervised learning .....	8
1.2.3 Reinforcement learning .....	11
1.2.4 The machine-learning toolbox .....	12
1.3 Basic terminology .....	13
1.3.1 Models .....	14
1.3.2 A closer look at what a model does★ .....	15
1.4 The machine learning workflow .....	17
<b>2 Data and attribute types</b> .....	19
2.1 What is a dataset? .....	19
2.1.1 Attributes .....	20
2.1.2 Attribute types .....	21
2.2 Data issues .....	22
2.3 The standard data format .....	23
2.4 Feature transformations .....	24
2.4.1 One-out-of-K coding .....	25
2.4.2 Binarizing/thresholding .....	26
Problems .....	27

<b>3</b>	<b>Principal Component Analysis</b> .....	29
3.1	Projections and subspaces★ .....	29
3.1.1	Subspaces .....	30
3.1.2	Projection onto a subspace .....	31
3.2	Principal Component Analysis .....	33
3.3	Singular Value Decomposition and PCA .....	39
3.3.1	The PCA algorithm .....	39
3.3.2	Variance explained by the PCA .....	40
3.4	Applications of principal component analysis .....	41
3.4.1	A high-dimensional example .....	42
3.4.2	Uses of PCA .....	44
	Problems .....	48
<b>4</b>	<b>Summary statistics and measures of similarity</b> .....	51
4.1	Attribute statistics .....	51
4.1.1	Covariance and Correlation .....	53
4.2	Term-document matrix .....	54
4.3	Measures of distance .....	55
4.3.1	The Mahalanobis Distance .....	57
4.4	Measures of similarity .....	57
	Problems .....	60
<b>5</b>	<b>Discrete probabilities and information</b> .....	61
5.1	Probability basics .....	62
5.1.1	A primer on binary propositions★ .....	63
5.1.2	Probabilities and plausibility .....	63
5.1.3	Basic rules of probability .....	65
5.1.4	Marginalization and Bayes' theorem .....	65
5.1.5	Mutually exclusive events .....	67
5.1.6	Equally likely events .....	68
5.2	Discrete data and stochastic variables .....	72
5.2.1	Example: Bayes theorem and the cars dataset .....	73
5.2.2	Generating random numbers★ .....	75
5.2.3	Expectations, mean and variance .....	76
5.3	Independence and conditional independence .....	77
5.4	The Bernoulli, categorical and binomial distributions .....	78
5.4.1	The Bernoulli distribution .....	78
5.4.2	The categorical distribution .....	79
5.4.3	Parameter transformations .....	80
5.4.4	Repeated events .....	80
5.4.5	A learning principle: Maximum likelihood .....	81
5.4.6	The binomial distribution★ .....	83
5.5	Information Theory★ .....	83
5.5.1	Measuring information .....	84
5.5.2	Entropy .....	86
5.5.3	Mutual information .....	87

- 5.5.4 Normalized mutual information ..... 88
- 6 Densities and models** ..... 91
  - 6.1 Probability densities ..... 91
    - 6.1.1 Multiple continuous parameters ..... 92
  - 6.2 Expectations, mean and variance ..... 95
  - 6.3 Examples of densities ..... 96
    - 6.3.1 The normal and multivariate normal distribution ..... 97
    - 6.3.2 Diagonal covariance ..... 99
    - 6.3.3 The Beta distribution ..... 100
    - 6.3.4 The central limit theorem★ ..... 101
  - 6.4 Bayesian probabilities and machine learning ..... 103
    - 6.4.1 Choosing the prior ..... 105
  - 6.5 Bayesian learning in general ..... 106
  - Problems ..... 109
- 7 Data Visualization** ..... 111
  - 7.1 Basic plotting ..... 111
  - 7.2 What sets apart a good plot? ..... 118
  - 7.3 Visualizing the machine-learning workflow★ ..... 119
    - 7.3.1 Visualizations to understand loss ..... 119
    - 7.3.2 Use visualizations to understand mistakes ..... 120
    - 7.3.3 Visualization to debug methods ..... 121
    - 7.3.4 Use visualization for an overview ..... 122
    - 7.3.5 Illustration of baseline and ceiling performance ..... 125
    - 7.3.6 Visualizing learning curves ..... 126
  - Problems ..... 128

---

**Part II Supervised learning**


---

<b>8</b>	<b>Introduction to classification and regression</b> .....	133
8.1	Linear models .....	133
8.1.1	Training the linear regression model .....	135
8.2	Logistic Regression .....	139
8.2.1	The confusion matrix .....	141
8.3	The general linear model★ .....	143
	Problems .....	145
<b>9</b>	<b>Tree-based methods</b> .....	147
9.1	Classification trees .....	148
9.1.1	Impurity measures and purity gains .....	148
9.1.2	Controlling tree complexity .....	152
9.2	Regression trees .....	154
	Problems .....	157
<b>10</b>	<b>Overfitting and performance evaluation</b> .....	159
10.1	Cross-validation .....	159
10.1.1	A simple example, linear regression .....	159
10.1.2	The basic setup for cross-validation .....	161
10.1.3	Cross-validation for quantifying generalization .....	164
10.1.4	Cross-validation for model selection .....	166
10.1.5	Two-layer cross-validation .....	166
10.2	Sequential feature selection .....	170
10.2.1	Forward Selection .....	171
10.2.2	Backward Selection .....	173
10.3	Cross validation of time-series data★ .....	173
10.3.1	The setup .....	174
10.3.2	Cross-validation .....	177
10.3.3	Two-layer cross-validation .....	178
10.4	Quantitative evaluation and comparison of classifiers .....	178
10.4.1	The general solution .....	180
10.4.2	First task: Evaluation of a single classifier★ .....	181
10.4.3	Second task: Comparing two classifiers★ .....	183
10.4.4	Comments★ .....	186
10.5	Visualizing learning curves★ .....	187
10.5.1	The setup .....	188
<b>11</b>	<b>Nearest neighbour methods</b> .....	191
11.1	K-nearest neighbour classification .....	191
11.1.1	A Bayesian view of the KNN classifier★ .....	192
11.2	K-nearest neighbour regression .....	195
11.2.1	Higher-order KNN regression★ .....	196
11.3	Cross-validation and nearest-neighbour methods .....	196

Problems ..... 199

**12 Bayesian methods** ..... 201

12.1 Discriminative and generative modelling ..... 201

12.1.1 Bayes classifier ..... 202

12.2 Naïve-Bayes classifier ..... 203

12.2.1 Robust estimation ..... 205

12.3 Bayesian networks ..... 206

12.3.1 A brief comment on causality ..... 210

Problems ..... 211

**13 Regularization and the bias-variance decomposition** ..... 213

13.1 Least squares regularization ..... 213

13.1.1 The effect of regularization ..... 214

13.2 Bias-variance decomposition ..... 217

Problems ..... 222

**14 Neural Networks** ..... 223

14.1 The feedforward neural network ..... 223

14.1.1 Artificial neural networks ..... 223

14.1.2 The forward pass in details ..... 224

14.2 Training neural networks ..... 227

14.2.1 Gradient Descent★ ..... 228

14.3 Neural networks for classification ..... 231

14.3.1 Neural networks for binary classification ..... 231

14.3.2 Neural networks for multi-class classification ..... 232

14.3.3 Multinomial regression ..... 233

14.3.4 Flexibility and cross-validation ..... 234

14.4 Advanced topics★ ..... 234

14.4.1 Mini-batching ..... 234

14.4.2 Convolutional neural networks ..... 235

14.4.3 Autoencoders ..... 236

14.4.4 Recurrent neural networks ..... 236

14.4.5 Serious neural network modelling ..... 237

Problems ..... 238

**15 Performance evaluation and class imbalance** ..... 241

15.1 Dealing with class imbalance ..... 241

15.1.1 Resampling ..... 242

15.1.2 Penalization ..... 242

15.2 Area-under-curve (AUC) ..... 244

15.2.1 The confusion matrix and thresholding ..... 245

Problems ..... 249

<b>16 Ensemble methods</b> .....	251
16.1 Introduction to ensemble methods .....	251
16.2 Bagging .....	253
16.3 Random Forests .....	255
16.4 Boosting .....	256
16.4.1 AdaBoost .....	257
16.4.2 Properties of the AdaBoost algorithm★ .....	260
Problems .....	262

---

**Part III Unsupervised learning**

---

<b>17 Distance-based clustering techniques</b> .....	265
17.1 Types of clusters .....	265
17.1.1 The distance-based cluster types .....	265
17.1.2 More elaborate cluster types .....	266
17.2 <i>K</i> -means clustering .....	266
17.2.1 A closer look at the <i>K</i> -means algorithm .....	269
17.2.2 Practical issues with the <i>K</i> -means algorithm .....	270
17.3 Hierarchical agglomerative clustering.....	271
17.3.1 Selecting linkage function.....	273
17.4 Comparing partitions .....	277
17.4.1 Rand index .....	280
17.4.2 Jaccard similarity .....	282
17.4.3 Comparing partitions using normalized mutual information .....	283
Problems .....	286
<b>18 Mixture models for unsupervised clustering</b> .....	289
18.1 The Gaussian mixture model.....	289
18.2 The EM algorithm.....	292
18.2.1 Why the EM algorithm works★ .....	295
18.2.2 Some problems with the EM algorithm.....	297
18.2.3 Selecting <i>K</i> for the GMM using Cross-validation .....	298
Problems .....	300
<b>19 Density estimation</b> .....	303
19.1 The kernel density estimator .....	303
19.1.1 Selecting the kernel width $\lambda$ .....	304
19.2 Average relative density .....	306
Problems .....	310
<b>20 Association rule learning</b> .....	313
20.1 Basic concepts .....	313
20.1.1 Itemsets and association rules .....	314
20.1.2 Support .....	315
20.1.3 Confidence .....	315
20.2 The Apriori algorithm .....	316
20.2.1 An example of the Apriori algorithm.....	317
20.3 Using the Apriori algorithm to find itemsets with high confidence.....	319
20.4 Some limitations .....	320
Problems .....	321
<b>Solutions</b> .....	323

XIV Contents

<b>A Mathematical Notation</b> .....	339
Elementary notation .....	340
Linear Algebra .....	340
Analysis .....	341
Probability Theory .....	343
<b>References</b> .....	345