

Master Machine Learning with Python

by
Ted Petrou

© 2019 Ted Petrou All Rights Reserved

Contents

I	Introduction to Machine Learning	1
1	Course Data	3
1.1	Introducing the Ames, Iowa Housing Dataset	3
1.2	The data dictionary	4
2	Learning vs Machine Learning	7
2.1	What is Learning?	7
2.2	What is Machine Learning?	7
2.3	The two types of machine learning	8
2.4	Terminology	9
3	The Machine Learning Model	11
3.1	What is a model?	11
3.2	Machine learning models	11
3.3	Create each model	12
3.4	Plotting a range of predictions	13
4	Assessing task performance	15
4.1	Assessing regression task performance	15
4.2	Comparing model performance	18
4.3	Exercises	24
5	Exploratory Data Analysis	25
5.1	Essential data information	25
5.2	Kinds of data	26
5.3	Univariate analysis	27
5.4	Multivariate analysis	39
5.5	Using pivot tables for three or more variables	42
5.6	Correlation to SalePrice	42
5.7	Continuing EDA	43
II	Linear Regression	45
6	Linear Regression	47
6.1	The linear regression model	47
6.2	Propose search area for w_0 and w_1	49
6.3	Plot the model along with the actual data	51
6.4	R-squared - a slightly different error metric	52
6.5	Exercises	56
7	Linear Regression Widget	57

8	Linear Regression with scikit-learn	59
8.1	The scikit-learn Estimator	59
8.2	The three-step process for each estimator - Import, Instantiate, Fit	60
8.3	What happens during fit?	63
8.4	Exercises	64
9	Prediction and Performance Evaluation	65
9.1	Repeating the three step machine learning process	65
9.2	Evaluating the performance of our predictions	67
9.3	Exercises	67
10	Multiple Linear Regression	69
10.1	Same Goal - minimize squared error	69
10.2	Choose features to build a model	69
10.3	Import, Instantiate, Fit	70
10.4	Make predictions	70
10.5	Evaluating model performance	71
10.6	Comparing multivariate and univariate model	72
10.7	Multiple linear regression model interpretation	72
10.8	Exercises	73
11	Establishing a Baseline	75
11.1	How to establish a baseline?	76
11.2	The dummy module	76
11.3	Exercises	76
12	What it means to be a linear model	79
12.1	Linear regression is more flexible than a straight line	80
12.2	Create new input data	81
12.3	Feature Engineering	82
12.4	Exercises	82
III	More Supervised Learning Models	85
13	K-Nearest Neighbors	87
13.1	How KNN works	87
13.2	KNN in pandas	87
13.3	KNN with multiple features	89
13.4	Visualizing KNN	90
13.5	Use Scikit-Learn	92
13.6	Measuring Performance of KNN	93
13.7	Notes on KNN	93
13.8	Distance calculation on features of different scale	94
13.9	The fewer the neighbors the higher the variance	94
13.10	Exercises	95
14	Decision Trees	97
14.1	How a decision tree is created	98
14.2	Step 6 - Repeat the above steps for each node until some stopping criterion is met	103
14.3	Use scikit-learn to create a decision tree	107
14.4	Exercises	109

15 Random Forests	111
15.1 Random forests are a collection of decision trees	111
15.2 Random Forest in Scikit-Learn	112
15.3 Random forests build weak learners, why are they good?	115
15.4 Exercises	115
IV Model Evaluation	117
16 Evaluating Model Performance	119
16.1 First idea - split data into a training and test set	119
16.2 Fit model just on the training data	121
16.3 Next Idea - Cross Validation	122
16.4 K-Fold Cross Validation in Scikit-Learn	122
16.5 Other flavors of cross validation	123
16.6 Cross validation on other models	124
16.7 No model is returned	125
16.8 Exercises	126
17 Evaluation Metrics	127
17.1 Different metrics with cross validation	129
17.2 Exercises	131
V Model Selection	133
18 Hyperparameter Tuning	135
18.1 Hyperparameters vs Parameters	135
18.2 Overfitting	135
18.3 Inspecting the decision tree	137
18.4 Change hyperparameters to reduce overfitting	138
18.5 Model selection	140
18.6 Optimizing other hyperparameters	141
18.7 Setting multiple hyperparameters simultaneously	143
18.8 Hyperparameter tuning with k-nearest neighbors	145
18.9 Hyperparameter tuning linear regression	147
18.10 Setting hyperparameters is like setting specifications for a car	147
18.11 Exercises	148
19 Automating Hyperparameter Tuning	149
19.1 The GridSearchCV meta-estimator	149
19.2 Grid searching multiple hyperparameters	152
19.3 Grid searching is computationally expensive	154
19.4 Reduce computation time with RandomizedSearchCV	155
19.5 Using different metrics when grid searching	158
19.6 Hyperparameter tuning is helpful but not the most important thing	159
19.7 Exercises	159
VI Data Transformations	161
20 Missing Value Imputation	163
20.1 Imputation in scikit-learn	164

20.2	Common mistake - filling with mean of new data	168
20.3	Summary of simple imputation	169
20.4	K-nearest neighbor imputation	169
20.5	Iterative imputation (experimental)	172
20.6	Exercises	173
21	Feature Scaling	175
21.1	Comparing numbers	175
21.2	Common feature scaling	175
21.3	Feature scaling in scikit-learn	176
21.4	Machine learning with scaled features	180
21.5	Linear regression coefficients using scaled data	184
21.6	Exercises	185
22	Simple Pipelines	187
22.1	Successive transformations without a Pipeline	187
22.2	Automating transformations with the Pipeline	188
22.3	Machine learning pipelines	189
22.4	Exercises	192
23	Grid Searching Pipelines	193
23.1	Using GridSearchCV on a pipeline	194
23.2	Exercises	195
24	Categorical Data	197
24.1	Encoding	198
24.2	Inverting the encoding	203
24.3	Ordinal Encoding	204
24.4	Machine learning with categorical data	205
24.5	Using features with different transformations	205
24.6	Exercises	206
25	The ColumnTransformer	207
25.1	Add transformation group to scale the continuous features	210
25.2	Machine learning after transforming	211
25.3	Create a pipeline within the ColumnTransformer	212
25.4	Grid searching the final pipeline	217
25.5	Exercises	217