Important points to remember

Machine learning is the systematic study of algorithms and systems that improve their knowledge or performance with experience. 3

Tasks are addressed by models, whereas learning problems are solved by learning algorithms that produce models. 12

Machine learning is concerned with using the right features to build the right models that achieve the right tasks. 12

Models lend the machine learning field diversity, but tasks and features give it unity. 13

Use likelihoods if you want to ignore the prior distribution or assume it uniform, and posterior probabilities otherwise. 28

Everything should be made as simple as possible, but not simpler. 30

In a coverage plot, classifiers with the same accuracy are connected by line segments with slope 1. 59

In a normalised coverage plot, line segments with slope 1 connect classifiers with the same average recall. 60

The area under the ROC curve is the ranking accuracy. 67

Grouping model ROC curves have as many line segments as there are instance space segments in the model; grading models have one line segment for each example in the data set. 69

https://doi.org/10.1017/CBO9780511973000.016 Published online by Cambridge University Press

By decreasing a model's refinement we sometimes achieve better ranking performance. 69

Concavities in ROC curves can be remedied by combining segments through tied scores. 77

To avoid overfitting, the number of parameters estimated from the data must be considerably less than the number of data points. 93

In descriptive learning the task and learning problem coincide. 95

The LGG is the most conservative generalisation that we can learn from the data. 108

Every concept between the least general one and one of the most general ones is also a possible hypothesis. 112

An upward path through the hypothesis space corresponds to a coverage curve. 114

Decision trees are strictly more expressive than conjunctive concepts. 131

One way to avoid overfitting and encourage learning is to deliberately choose a restrictive hypothesis language. 131

The ranking obtained from the empirical probabilities in the leaves of a decision tree yields a convex ROC curve on the training data. 138

Entropy and Gini index are sensitive to fluctuations in the class distribution, $\sqrt{\text{Gini}}$ isn't. 147

Rule lists are similar to decision trees in that the empirical probabilities associated with each rule yield convex ROC and coverage curves on the training data. 166

$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ acts as a transformation that decorrelates, centres and normalises the features. 202

Assuming uncorrelated features effectively decomposes a multivariate regression problem into d univariate problems. 203

A general way of constructing a linear classifier with decision boundary $\mathbf{w}\cdot\mathbf{x} = t$ is by constructing $\mathbf{w}$ as $\mathbf{M}^{-1}(n^{\oplus}\boldsymbol{\mu}^{\oplus} - n^{\ominus}\boldsymbol{\mu}^{\ominus})$. 207

In the dual, instance-based view of linear classification we are learning instance weights α_i rather than feature weights w_j. 210

A minimal-complexity soft margin classifier summarises the classes by their class means in a way very similar to the basic linear classifier. 219

The basic linear classifier can be interpreted from a distance-based perspective as constructing exemplars that minimise squared Euclidean distance within each class, and then applying a nearest-exemplar decision rule. 239

https://doi.org/10.1017/CBO9780511973000.016 Published online by Cambridge University Press

Probabilities do not have to be interpreted as estimates of relative frequencies, but can carry the more general meaning of (possibly subjective) degrees of belief. 265

For uncorrelated, unit-variance Gaussian features, the basic linear classifier is Bayes-optimal. 271

The negative logarithm of the Gaussian likelihood can be interpreted as a squared distance. 271

A good probabilistic treatment of a machine learning problem achieves a balance between solid theoretical foundations and the pragmatism required to obtain a workable solution. 273

An often overlooked consequence of having uncalibrated probability estimates such as those produced by naive Bayes is that both the ML and MAP decision rules become inadequate. 277

Tree models ignore the scale of quantitative features, treating them as ordinal. 305

Fitting data to a fixed linear decision boundary in log-odds space by means of feature calibration can be understood as training a naive Bayes model. 318

Low-bias models tend to have high variance, and vice versa. 338

Bagging is predominantly a variance-reduction technique, while boosting is primarily a bias-reduction technique. 339

Machine learning experiments pose questions about models that we try to answer by means of measurements on data. 343

The combination of precision and recall, and therefore the F-measure, is insensitive to the number of true negatives. 346

Confidence intervals are statements about estimates rather than statements about the true value of the evaluation measure. 352

https://doi.org/10.1017/CBO9780511973000.016 Published online by Cambridge University Press

https://doi.org/10.1017/CBO9780511973000.016 Published online by Cambridge University Press