4/25/2022

Classification

Grouping the data into distinct classes or categories (categorical or discrete).
Binary Classifier, Multi-class classifiers, Multi-label classifiers

Lazy learner vs. eager learner

Lazy learner predicts directly from the training data itself
Eager learner will create a model, and then base predictions on that model (such as regression)

Logistic regression – binary classifier

Naïve Bayes classifier: needs very little data. Fast. But unfortunately it's a poor predictor. (based on a two-way table/crosstabs, and assumes independence.)

Stochastic gradient descent – many hyperparameters, it can be sensitive to scaling, depends on the derivative to find the best "descent" path to find the minimum error.

KNN – K nearest neighbors – finds the shortest distance to a value (or k values) in the training set, and then assigns the test value the same category (voting is applied, majority wins). It's best to use odd number of nearest neighbors, because ties are randomly decided.

Decision Trees – can be quite unstable. It is generally overfitted. (test data never does as well as training data) – bagging

Random Forest – ensemble method; builds a number of small trees, using random combinations of variables, and then "votes" on which class the predicted value belongs to. – boosting

Neural Network – high noise tolerance, hard to interpret

SVM – support vector machines – memory efficient, useful for high dimensional data. The simplest form uses a line to divide the categories (a plane, hyperplane). The plane/line is in the middle of all the data. That line is placed so that the mistakes are as few as possible. The optimal separation line. Can project onto higher dimensions to obtain non-linear separators. (also a binary classifier)

Evaluate classifiers:
Test/train split – hold-out method
Cross validation – k-fold cross validation
Accuracy – measured as a percent of correctly classified values
F-1 score (weighted average of precision and recall)
ROC curve

General methods:

- Read the data
- Create dependent and independent data sets
- Split into test/train
- Train the model with different classifiers

- Choose the best classifiers with the most accuracy

Regression

Linear Regression – ordinary least squares

Nonlinear Regression – polynomials models, splines, log models, power models, etc. and can include interaction terms.

Model Selection:

Stepwise method
Best Subset selection
PCA – principal component analysis
PLA – principal least squares
Penalized Regression – LASSO, Ridge regression, etc.

Metrics:

Root mean square error (RMSE)
Adjusted R^2
K-fold cross validation

Graph Theory:

Deep learning – neural networks
Tree-based models
Markov chain models

Python –

Sklearn (sci-kit learn), contains most major machine learning tools
Keras tensor flows, pytorch for neural networks

Data Equity

Especially when dealing with classification models, where one group or class is significantly smaller than the other classes. This is called masking. It may be more efficient (accuracy) to overlook the smaller class and mispredict that class in order to get higher accuracy on the more dominant class. You may get high overall accuracy, but the small class may be consistently mispredicted.

This is problematic, especially when dealing with people and groups of people that are minorities, or are systematically underrepresented. The models can then reproduce the biases of the culture that the data was drawn from.

When the classes are of unequal sizes, you may need to adjust the class sizes in the training data to produce more equally sized input groups. Spread the inaccuracy over all the groups instead of just the one.

Sentiment Analysis:

A type of NLP, to detect positive or negative sentiment, other feelings like anger, or urgency, intention or interest

Mutli-lingual sentiment analysis – not very good at switching between languages. Now use language detection tool to identify the language, then use a single-language sentiment analysis method.