4/19/2021

Classification

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

Lazy learner vs. eager learner

Lazy learner predicts directly from the training set itself.
Eager learner will create a model, and then predictions are based on the model (similar to regression)

Logistic regression – binary classifier

Naïve Bayes classifier. Needs very little data. Fast. Unfortunately: also a poor predictor. (classes are independent).

Stochastic gradient descent – many hyperparameters, sensitive to scaling, depends on 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; ties are decided randomly

Decision Trees – is quite unstable, is generally overfitted (test data won’t do as well as training). – bagging

Random Forest – ensemble method : builds a number of small trees, using random combinations of variable, and then “votes” on which class the predicted value belongs in -- boosting

Neural network – high noise tolerance, hard to interpret

SVM – support vector machines – memory efficient, useful for high dimensional spaces. The simplest form uses a line to divide the categories (in the plane) – a hyperplane in higher dimensional spaces. The plane/line is the line that is in the middle of the all data—the optimal separation line. Can project into higher dimensional spaces to obtain non-linear separators. (also a binary classifier).

Evaluating 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 model with different classifiers

- Evaluate the classifier
- Choose the best classifier with most accuracy

Regression

Linear regression – ordinary least squares
Nonlinear regression – polynomial models, spline, log models, power models, etc. interaction terms, etc.

Model Selection:
Stepwise method
Best Subset selection
PCA – partial least squares
Penalized Regression – LASSO, Ridge regression

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 – scikit learn, most major/common machine learning models
Keras tensor flows, pytorch with network models

Data Equity –
Especially when dealing with classification models, where one group or class is significantly smaller than the other(s) class(es). This is called "masking". It may be more efficient (accurate) to overlook and mis-predict the smallest class in order to improve accuracy on the larger class. You may get high overall accuracy, but the smaller class will constantly be 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 produced it.

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

Sentiment Analysis:
A type of NLP, to detect positive or negative sentiment, other feelings (like anger), urgency, intention/interest

Multi-lingual sentiment analysis – not very good at switching between languages. Now use language detection tools to identify the language, and then use single language sentiment analysis methods.