Machine Learning Models (Tom Springett)

# Supervised Learning

Finding Donors

List three of the supervised learning models above that are appropriate for this problem that you will test on the census data. For each model chosen

- Describe one real-world application in industry where the model can be applied.

- What are the strengths of the model; when does it perform well?

- What are the weaknesses of the model; when does it perform poorly?

- What makes this model a good candidate for the problem, given what you know about the data?

##### KNN

* Can be used for both regression and classification, but more often for classification.
* Good when data is clustered.

#### Applications

* Animal recognition
* Marketing campaign to identify customers
* Face Recognition

#### Strengths

* Simple and easy to implement
* No need to tune several parameters

#### Weaknesses

* Gets significantly slower as the number of features increase
* May require dimension reduction and PCA

#### Why is it a good Candidate?

* Small dataset.
* Easy to understand

#### References

* Machine Learning Basics with K-Nearest Neighbors. Medium.
* Intro to K-Nearest Neighbors. Analytics Vidhya.
* K-Nearest Neighbors. Wikipedia.

##### SVM

A close up of a map

Description automatically generated

* The algorithm creates a line or a hyperplane which separates the data into classes
* Separates out classes. Data points considered as vectors. Best hyperplane has most margin.
* Support vectors are the data points closest to the opposing class
* Data may be transformed into higher dimensional space for better separation
* Kernel trick may be used to reduce computation. Dot product used to transform non-linear to linear space.

#### Applications

* Home mortgage approval
* Medical imaging
* Study air quality
* Financial analysis

#### Strengths

* Effective in high dimensional spaces. When the number of features is large, for example text classivation in
* Uses a subset of training points in the decision function, support vectors.

#### Weaknesses

* Not good on imbalanced or skewed dataset. For example, customer terms with low DPM (defects per million). Logistic Regeression is better on these datasets.
* The complex data transformations can be computationally slow.
* Gets significantly slower as the number of features increase
* Poor performance when number of features is greater than number of samples
* A number of parameters to optimize (polynomial or linear kernel, Gamma (how far away from separation line points are considered, C or regularization parameter).
  + A large value of C means a number of training points will be considered, but there may be overfitting.
  + Gamma defines how far a single data point reaches. A low gamma even the far away point get considered and there is less over-fitting.

#### Why is it a good Candidate?

* Small dataset
* Has a good margin of separation
* Number of samples is greater than number of features

#### References

* Understanding Support Vector Machine algorithm from examples (along with code). Analytics Vidhya.
* Support Vector Machines Fun and Easy. Youtube Video
* Support Vector Machines. Towards Data Science

##### Gaussian Naïve Bayes

* Classification algorithm based on probabilities. They are calculated for every feature and the ones with the highest probability are selected.
* The prior probability and posterior (updated probability of B given A) probability are key terms. For example, for 3 lots and one has oil below it, the prior probability that C has oil is 30%. If a well is drilled on lot B, the posterior probability that lot C has oil is 50%. The posterior probability takes the new information into account.

#### Applications

* Algorithm of choice for text characterization
* Housing prices

#### Strengths

* Fast. Only the mean and standard deviation probabilities need to be calculated.
* Good with complex situation. Only a small training data set is required.

#### Weaknesses

* Assumes conditional independence

#### Why is it a good Candidate?

* Small

#### References

* Naïve Bayes in Machine Learning. Medium.
* Investopedia

##### Decision Tree

* Predominantly used for classifying subjects into known groups. Can be used for both classification and regression.

#### Applications

* Knowledge management platforms for customer service
* Customer wiliness to buy a product
* In finance, forecasting future outcomes and probabilities
* Loan Approval/
* How to set housing prices

#### Strengths

* Simple to understand

#### Weaknesses

* Over fitting
* Can be unstable with small variation. Completing different trees can be easily generated. Bagging and Boosting can help.
* Need to determine when to stop splitting
* When to set maximum depth
* Pruning or removal of branches to reduce overfitting.

#### Why is it a good Candidate?

* Small

#### References

* When to use Linear Regression, Clustering or Decision Trees. Dzone.
* Decision trees in real learning. Towards data science (Medium)

##### Template

* Predominantly used for classification.

#### Applications

* H

#### Strengths

* EffectWeaknesses
* Gets

#### Weaknesses

#### Why is it a good Candidate?

* Small

#### References

* Understanding

Here is a summary from scikit learn (Choosing the right estimator):

A screenshot of a cell phone

Description automatically generated