Features Selection

Week 05 - Day 03

Questions we want to answer

What are the useless features?

What is the best

combination of features?

Solution: features selection!

Naive Approach

Try all possible combinations

With 10 features -> 1000 combinations

With 40 features -> ???

With 10 features -> 1000 combinations

With 40 features -> 1,100,000,000,000

Pure Optimization

Simulated Annealing

Genetic Algorithms

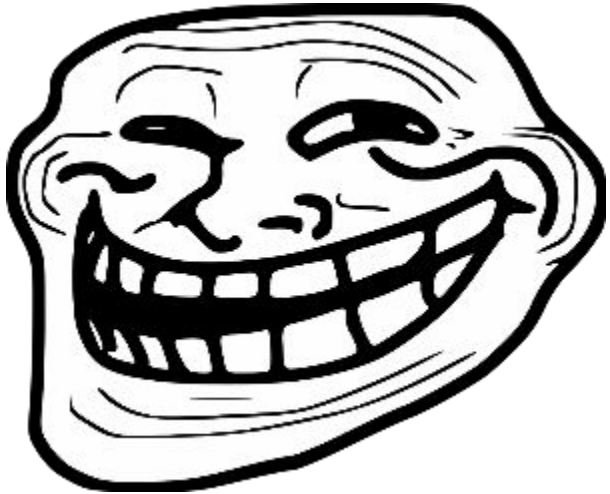
Hill Climbing

Particle Swarm Optimization

Black Mamba Optimization

Black Mamba Optimization

doesn’t exist



Stupid moment of the day:

Is it pokemon or big data?

Simple approach 1: Bottom-up Approach

1. Start with an empty set
2. Add features one by one
3. Stop if the model is not improving

Choose a metric to insert a new feature

(e.g. correlation coefficient)

Cons:

a. doesn’t consider interaction between features
b. Once the feature is in, it cannot be removed

Simple approach 2: Top-down Approach

1. Start with all the features
2. Remove the features one by one
3. Stop if the model not improving

Choose how to remove a features

(e.g. smallest coefficient)

Cons:

a. Computationally expensive

Mixed approach

1. Start with empty set
2. Add p features
3. Remove q features

Random Shuffling

1. Create a model
2. Randomly shuffle a column
3. Check the score of the new model
4. Drop the column if we see no changes

Regularization

Loss function = error + penalty

Lasso regularization

Regularization works

also for logistic regression

(sklearn = “penalty”)

Other simple feature selection techniques

Remove features with low variance

NLP: remove words that appears in less than x% of the documents

Sklearn

http://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_selection

http://scikit-learn.org/stable/modules/feature_selection.html#univariate-feature-selection