KNN Implementation

Chen Yu
UT Austin

Parameter selection: How to determine K?

- The goal is to produce correct answers on unseen instances
- During training, given training set (x1,y1), (x2,y2), …, (xn,yn). We write kNN code. Now we have a classifier that can predict the output category based on a new input Xnew.
- Since we don't know what new cases we will get, the only way to determine k is to use training data, more specially, measuring training set accuracy.

Parameter selection by training data

During training, given training set (x1,y1), (x2,y2), …, (xn,yn). We write kNN code. Now we have a classifier that can predict the output category based on a new input X_{new}.

We try different values of K, 1, 3, 5, 9…

We measure the accuracy on training examples in each case.

We select K that maximizes the predicts on training data

Overfitting or overgeneralization

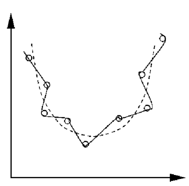
It does an excellent job of fitting the training data points

Overfitting is when a learning algorithm performs too good on the training set, compared to its true performance on unseen testing data.

Never use training accuracy to select parameters.

It does not reflect the structure which we expect to be present in unseen data. Instead, overfitting also fits noise in training data, not the general underlying regularity.

Overfitting

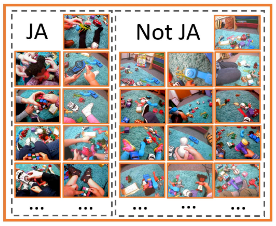


One big theoretical question in machine learning is how to get good generalization with a limited number of samples.

Building a ML system

- Data collection
- Data exploration: get familiar with data and understand the data so you can make informed decisions during the following steps.
  e.g. descriptive stats, visualization, identifying outliers and missing values.
- Data cleaning: remove outlier and noisy data points
- Preprocessing: reformatting the data
- Training and model evaluation: e.g. fine-tune parameters
- Interpretation of results

Joint Attention from egocentric vision



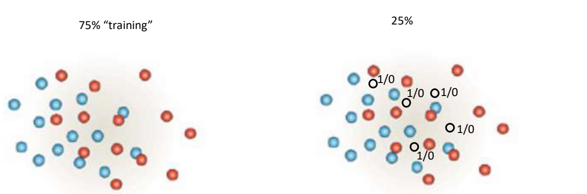
73

Splitting labeled data into a training set (e.g. 75%) and a testing set (e.g. 25%)

- Using the training set to train the model
- Using the test set to evaluate model performance.
- If the trained model performs above chance, then we can conclude that the information in the training set allows the model to distinguish the two classes. Therefore, we can further conclude that there are social signals in the data that can be potentially used to detect joint attention.

74

How to implement KNNs



Predicted results: 1110100100100111000000000000100011
Ground truth: 1010000011100011000000000011010011

Compare the results from KNNs with ground truth

75

Evaluation

- The classification accuracy is 70%. Is it good enough to confirm our hypothesis?
  NO!
- Unbalanced classes
- How can we evaluate performance in a more reasonable way?

Predicted results: 1110100100100111000000000000100011
Ground truth: 1010000011100011000000000011010011

76

Confusion matrix, precision and recall

	11 (true positive, TP)	10 (false positive, FP)	Model: 1; ground truth: 0
Model: 0; ground truth: 1	01 (false negative, FN)	00 (tru negative, TN)	

Precision = TP/(TP+FP), the ratio of correct positives among all positives predicted by the classifier.

Recall = TP/(TP+FN), the ratio of positive instances that are correctly detected by the classifier.

77

Confusion matrix, precision and recall

Taking a shot and hitting it	Taking a shot but missing it
Missing an opportunity that you could take a shot	

Precision = TP/(TP+FP), how precise you are

Recall = TP/(TP+FN), how many you hit

78

Precision and Recall

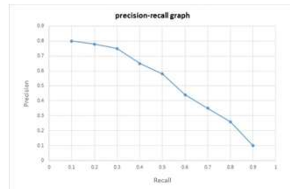
Precision = TP/(TP+FP), how precise you are

Recall = TP/(TP+FN), how many you hit

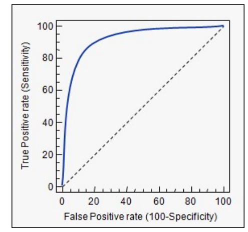
Ground truth:	1010000011	precision	recall
Aggressive:	1111111111	4/10=40%	100%
Conservative:	1000000000	1/1=100%	1/4=25%

79

Precision/Recall Tradeoff and Receiver Operating Characteristic (ROC)



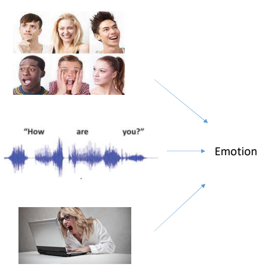
recall

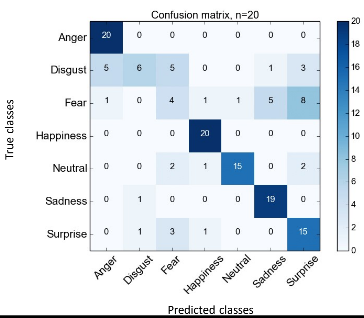


The ratio of negative instances that are incorrectly classified as positive

80

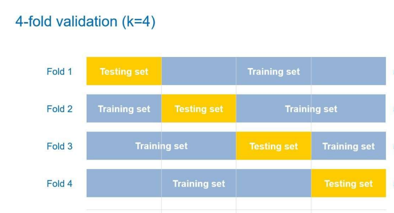
Multiclass classification





81

Cross validation



82

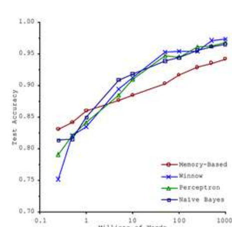
Main Challenges

- Nonrepresentative Training data
- Poor-Quality data
- Irrelevant Features

83

Main Challenges

- Insufficient quantity of training data



Banko, M. and Brill, E. (2001) , "Scaling to Very Very Large Corpora for Natural Language Disambiguation"

84