KYOTO UNIVERSITY

Evaluating Your Machine Learning Models

Hisashi Kashima
kashima@i.Kyoto-u.ac.jp

DEPARTMENT OF INTELLIGENCE SCIENCE AND TECHNOLOGY

Topics:
Performance measures and evaluation frameworks

- You want to know the final performance of your model, or select the best one among possible models (or both)
- Performance measure: accuracy, precision/recall, DCG@k, AUC
- Evaluation framework: cross validation

Performance Measures

Various performance measures: Should be chosen according to your applications

- There are various evaluation measure to quantify the performance of a trained model especially in supervised learning
  - Accuracy, precision/recall, DCG@k, AUC, …
- They should be appropriately chosen depending on applications
  - Classification with decision thresholds: accuracy, precision/recall, …
  - Classification without decision thresholds: AUC, …
  - Ranking: DCG@k, ...

Decision model and confusion matrix:
Decisions on a dataset give a confusion matrix

- The trained model gives confidence $P(\boldsymbol{x})$ on given instance $\boldsymbol{x}$ belonging to the positive class $(+1)$
  - Multi-class case: 1-vs-rest
- Assign $+1$ to $\boldsymbol{x}$ whose $P(\boldsymbol{x})$ is larger then decision threshold τ
- Fixing a model, a dataset, and a decision threshold gives a confusion matrix

		predicted label	
		positive	negative
true label	positive	#true positives ☺	#false negatives
	negative	#false positives	#true negatives ☺

Accuracy and precision/recall: Basic predictive performance measures

- Accuracy: percentage of #true_positives + #true_negatives
- Precision/Recall
  - Precision: #true_positives / (#true_positives+#false_positive)
  - Recall: #true_positives / (#true_positives+#false_negatives)
  - F-measure: Precision・Recall /(precision+recall)
    - an integrated measure of precision and recall

		predicted label	
		positive	negative
true label	positive	#true positives ☺	#false negatives
	negative	#false positives	#true negatives ☺

DCG@k: Performance measure for ranking

- In ranking (of web pages), accuracy of top-ranked items is more important
- Precision@k: precision calculated using the top-k scored items
- DCG(Discounted Cumulative Gain)@k is a weighted variant of Precision@k: $\sum_{i=1}^{k} \frac{rel(i)}{\log(i+1)}$

  $-rel(i)$ is the relevance score for the i-th ranked item

AUC:
Performance measure not depending on the threshold

- Evaluation needs fixing the decision threshold
- Imbalanced data generally results in a high accuracy
- AUC:
  - Performance measure directly work with confidence score $P(\boldsymbol{x})$
  - Probability of A being larger than B
    - A: confidence score of a randomly chosen positive instance
    - B: confidence score of a randomly chosen positive instance
  - takes 1 for perfect predictions, 0.5 for random predictions

Evaluation Framework

Evaluation framework:
We want to predict model performance

- Performance for the training data and that for the test data are different
  - What we are interested in is the latter
- Many models have hyper-parameters to be specified by users

First principle:
Evaluation must use a dataset not used in training

- You must not evaluate your classifier on the dataset you used for training
- Usually, first divide a given dataset into a training dataset and a test dataset
  - Train a classifier using the training dataset
  - Evaluate its performance on the test dataset
- Sometimes ordering of data instances (unintentionally) has some patterns in their labels
  - Partitioning should be done carefully

Cross validation (for performance testing): A statistical framework for performance evaluation

- You want to know the performance of the classifier (will be obtained using your algorithm) when it is deployed
- (K-fold) cross validation do this
- Divide a given dataset into K non-overlapping sets
  - Use K-1 of them for training
  - Use the remaining one for testing
- Changing the "test" dataset K gives K measurements
  - Take their average to get a final performance measure

Cross validation for tuning hyper-parameters: A statistical framework for performance evaluation

- Most of machine learning algorithms have hyper-parameters
  - Hyper-parameters: Parameters not automatically tuned in the training phase; given by users
- (K-fold) cross validation can be used for this
  - Use K-1 of K sets for training models for various hyper-parameter settings
  - Use the remaining one for testing
  - Choose the hyper-parameter setting with the best averaged performance

Double loop of cross validation: Tuning hyper-parameters and performance evaluation

- Sometimes you want to do both hyper-parameter tuning and performance evaluation
- Doing both with one K-fold cross validation is guilty
  - You see the test for tuning hyper-parameters
- Double loop cross validation
  - Outer loop for performance evaluation
  - Inner loop for hyper-parameter tuning
  - High computational costs…

A simple alternative of double-loop cross validation: "Development set" approach

- A simple alternative for the double-loop cross validation
- "Development set" approach
  - Use K-2 of K sets for training
  - Use one for tuning hyper-parameters
  - Use one for testing