

Ensemble Methods

Knowledge Discovery and Data Mining 2 (VU) (707.004)

Roman Kern

KTI, TU Graz

2018-03-08

Outline

- 1 Introduction
- 2 Classification
- 3 Clustering

Introduction to Ensemble Methods

Motivation & Basics

Quick facts

- Basic Idea: Have **multiple models** and a **method to combine** them into a single one.
- Predominately used in classification and prediction
- Sometimes called: combined models, meta learning, committee machines, multiple classifier systems
- Ensemble methods do have a long history and used in statistics for more than 200 years

Types of ensembles

- ... different hypothesis
- ... different algorithms
- ... different parts of the data set

Motivation

- ... as every model has its limitations
- Goal: combine the strength of all models
- Improve the accuracy of using an ensemble
- Be more robust in regard to noise

Basic Approaches

- Averaging
- Voting
- Probabilistic methods

Combination of Models

- Need a function to combine the results from the models
- For real values output
 - ▶ Linear combination
 - ▶ Product rule
- For categorical output, e.g. class labels
 - ▶ Majority vote

Linear combination

- Simple form of combining the output of an ensemble
- Given T models, $f_t(y|x)$
- $g(y|x) = \sum_{t=1}^T w_t f_t(y|x)$
- Problem of estimating the optimal weights (w_t)
- Simple solution: use the uniform distribution: $w_t = 1/T$

Product rule

- Alternative form of combining the output of an ensemble
- $g(y|x) = \frac{1}{Z} \prod_{t=1}^T f_t(y|x)^{w_t}$
- ... where Z is a normalisation factor
- Again, estimating the weights is non-trivial

Majority Vote

- Combining the output, if categorical
- The models produce a label as output, e.g. $h_t(x) \in \{+1, -1\}$
- $H(x) = \text{sign}(\sum_{t=1}^T w_t h_t(x))$
- If the weights are non-uniform, it is a **weighted vote**

Selection of models

- The models should not be identical, i.e. produce identical results
- ... therefore an ensemble should represent a **degree of diversity**
- Two basic types of achieving this diversity
 - ▶ *Implicitly*, e.g. by integrating randomness (bagging)
 - ▶ *Explicitly*, e.g. integrate variance into the process (boosting)
- Most of the methods implicitly integrate diversity

Motivation for ensemble methods

- Statistical
 - ▶ Large number of hypothesis (in relation to training data-set)
 - ▶ Not clear, which hypothesis is the best
 - ▶ Using an ensemble reduces the risk of picking a bad model
- Computational
 - ▶ Avoid local minima
 - ▶ Partially addressed by heuristics
- Representational
 - ▶ A single model/hypothesis might not be able to represent the data

Dietterich, T. G. (2000). Ensemble methods in machine learning. In Multiple classifier systems (pp. 1-15).

Classification

Ensemble Methods for Classification

Underlying question

How much of the ensemble prediction is due to the accuracies of the individual models and how much due to their combination?

→ express the ensemble error as two terms:

- Error of individual models
- Impact of interactions, the **diversity**

Note: It depends on the combination, whether one can separate the two terms

Regression error for the linear combination

- Squared error of the ensemble regression
- $(g(x) - d)^2 = \frac{1}{T} \sum_{t=1}^T (g_t(x) - d)^2 - \frac{1}{T} \sum_{t=1}^T (g_t(x) - g(x))^2$
- First term: error of the individual models
- Second term: interactions between the predictions
- ... the ambiguity, ≥ 0
- \rightarrow Therefore it is preferable to increase the ambiguity (diversity)
- Smallprint: Actually there is a tradeoff of bias, variance and covariance, known as accuracy-diversity dilemma

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross-validation and active learning. In *Advances in neural information processing systems* (pp. 231–238). Cambridge, MA: MIT Press.
Kuncheva,

Classification error for the linear combination

- For a simple averaging ensemble (and some assumptions)
- $e_{ave} = e_{add} \left(\frac{1 + \delta(T-1)}{T} \right)$
- ... where e_{add} is the error of the individual model
- ... and δ being the correlation between the models

Tumer, K., & Ghosh, J. (1996). Error correlation and error reduction in ensemble classifiers. *Connection Science* 8(3-4), 385-403.

Basic Approaches

- Bagging - combines strong learners → reduce variance
- Boosting - combines weak learners → reduce bias
- Many more: mixture of experts, cascades, ...

Bootstrap Sampling

- Create a distribution of data-sets from a single data-set
- If used within ensemble methods, it is typically called **Bagging**
- Simple approach, but has proven to increase performance

Davison, A. C., & Hinkley, D. (2006). Bootstrap methods and their applications (8th ed.). Cambridge: Cambridge Series in Statistical and Probabilistic Mathematics

Bagging

- Each member of the ensemble is generated by a different data-set
- Good for **unstable models**
 - ▶ ... where small differences in the input data-set yield big differences in output
 - ▶ Also known as *high variance* models
- → not so good for simple models

Note: Bagging is an abbreviation for bootstrap aggregating

Breiman, L. (1998). Arcing classifiers. *Annals of Statistics*, 26(3), 801–845.

Bagging Algorithm (train)

- 1 Input: Ensemble size T , training set $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$
- 2 For each model M_t
 - 1 For n' times, where $n' \leq n$
 - 1 Sampling (random) from D with replacement
 - 2 Train model M_t with subset

Bagging Algorithm (classify)

- For classification typically majority vote
- For regression typically linear combination

Note: Subset may contain duplications, i.e. if $n' = n$

Boosting

- Family of ensemble learners
- Boost weak learners to a strong learner
- **Adaboost** is the most prominent one
- Weak learners need to be better than random guessing

Adaboost

- Basic idea: Weight the individual instances of the data-set
- Iteratively learn models and record their errors
- Distribute the effort of the next round on the mis-classified examples

Adaboost (train)

- 1 Input: Ensemble size T , training set $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$
- 2 Define a uniform distribution W_t over elements of D
- 3 For each model M_i
 - 1 Train model M_i using distribution W_t
 - 2 Calculate the error of model ϵ_t and weight $\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$
 - 3 ... if $\epsilon_t > 0.5$ break (and discard model)
 - 4 ... else update the distribution W_t according to ϵ_t

Adaboost (classify)

- Linear combination, $H(x) = \text{sign}\left(\sum_{t=1}^T \alpha_t h_t(x)\right)$

Stacked generalisation

- Basic idea: Have the output of a layer of classifiers as input to another layer
- For 2 layers:
 - 1 Split the training data-set into two parts
 - 2 Learn the first layer using the first part
 - 3 Classify the second part and
 - 4 ... take the decision as input for the second part

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks* 5(2), 241–259

Mixture of Experts

- Basic idea: some models should specialise on parts of the input space
- Ingredients
 - ▶ Base models (e.g. specialised models - so called experts)
 - ▶ Component to estimate probabilities, often called a gating network
- The gating networks learns to select the appropriate expert for parts of the input space

Mixture of Experts - Example #1

- Ensemble of base learners being combined using weighted linear combination
- The weight is found via a neural network
 - ▶ The neural network is learnt via the same input data-set

Mixture of Experts - Example #2

- Mixture of expert models are called mixture models
- e.g. the Expectations-maximisation algorithm

Cascade of classifiers

- Setting
 - ▶ Have a sequence of models, each with high hitrate ($\geq h$) and low false alarm rate ($< f$)
 - ▶ ... with increasing complexity
 - ▶ In the data-set the negative examples are more common
- The cascade is learnt via boosting
- For example:
 - ▶ For $h = 0.99$ and $f = 0.3$ and a cascade of size 10
 - ▶ ... one gets the hitrate of about 0.9 and a false alarm rate of about 0.000006

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In Computer Vision and Pattern Recognition, 2001.

Decision Stump

- Decision stumps are a popular choice for (some) ensemble learning
- ... as they are fast
- ... as they are less prone to overfitting
- A decision stump is a decision tree that only uses a single feature (attribute)

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11, 63–91.

Random Subset Method

- Basic idea: Instead of taking a subset of the data-set, use a subset of the feature set
- ... will work best, if there are many features
- ... and will not work as well if most of the features are just noise

Random Forest

- Combines two randomization strategies
 - ▶ Select random subset of the data-set to learn decision tree (bagging), e.g. select $n = 100$ random trees
 - ▶ Select random subset of features, e.g. select \sqrt{m} features
- Random forests are used to estimate the importance of features (by comparing the error using a feature vs. not using a feature)
- Typically good performance

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Multiclass Classification

- Basic idea: split a multi-class problem into a set binary classification problems
- e.g. Error correcting output codes

Kong, E. B., & Dietterich, T. G. (1995). Error-correcting output coding corrects bias and variance. In International conference on machine learning.

Ensemble classification for multi-class problems

- Have different base classifiers for different parts of the feature set
- Train all base classifiers using the training data-set
- Record their performance with cross-evaluation for each class
- ... have two thresholds, $min_{precision}$ and min_{recall}
- If the precision for a certain class and model is $\geq min_{precision} \rightarrow$ allowed to vote
- If the recall for a certain class and model is $\geq min_{recall} \rightarrow$ allowed to vote against (veto)
- In the classification use a weighted vote
 - ▶ where veto is a negative vote
 - ▶ ... and the weight is according to the respective measure (precision or recall)

Kern, R., Seifert, C., Zechner, M., & Granitzer, M. (2011, September). Vote/Veto Meta-Classifier for Authorship Identification Notebook for PAN at CLEF 2011.

- **Active learning** is a form of semi-supervised learning
- The basic idea is to give the human instances to label
- ... which carry the most information (to update the model)
- **Query by Committee**
- ... use an ensemble, i.e. the disagreement of multiple classifiers to pick instances

Clustering

... and other approaches

Cluster Ensembles

- Basic idea: Have multiple clustering algorithms group a data-set
- ... combine all results into a single clustering results
- Motivation: More reliable result than individual cluster solutions

Consensus Clustering

- Have a set of clusterings: $\{C_1, \dots, C_m\}$
- Find an overall clustering solution C
- Minimise the disagreement using a metric: $D(C) = \sum_{C_i} d(C, C_i)$
- Also known as clustering aggregation

Mirkin Metric

- The metric reflects the numbers of pairs of instances ...
- ... being together in the overall clustering, but separate in C_i
- ... and vice versa

- Ensemble methods are not limited to machine learning tasks alone
- For example, in the field of recommender systems they are known as **hybrid recommender system**
 - ▶ e.g. combine a content based recommender with a collaborative filtering one

The End

Next: Text Mining + Tools