Ensemble learning

Natural Computing

Gijs van Tulder
April 2020

Based on an earlier version by Jesse Krijthe.

Ensemble learning – 1/4 Introduction

Natural Computing

Gijs van Tulder
April 2020

Based on an earlier version by Jesse Krijthe.

Ensemble learning

Combine the output from different prediction models.

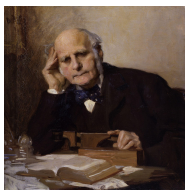
Ensemble learning

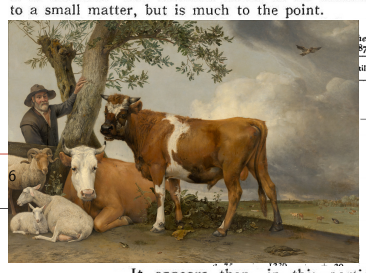
1. Introduction
2. How to combine predictions?
3. How to create different models
4. Boosting

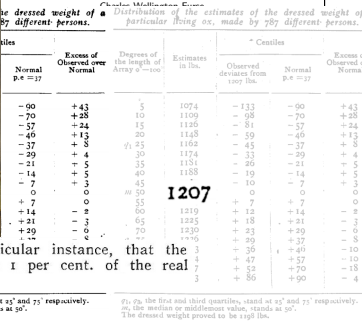
Combining airway extraction models

Lo et al. (2012). "Extraction of Airways From CT (EXACT'09)". Transactions on Medical Imaging

Sir Francis Galton (1822–1911)







VOX POPULI.

IN these democratic days, any investigation into the trustworthiness and peculiarities of popular judgments is of interest. The material about to be discussed refers to a small matter, but is much to the point.

1207

It appears then, in this particular instance, that the *vox populi* is correct to within 1 per cent. of the real value,

Mauritshuis, Den Haag

Ensemble learning

1. Obtain independent predictions from multiple predictors.
2. Combine the individual predictions.
3. Use the average prediction as the final result.

Ensemble learning

Why does this work?

What are the assumptions?

Condorcet's jury theorem

Twelve Angry Men (1957)

Marquis de Condorcet
(1743–1794)

When does it make sense to combine decisions?

If $p > 0.5$, adding voters to the jury increases the probability of a correct majority vote (assuming voting is independent).

Convergence of the mean

Single:

$$\mathrm{Var}\left(X_i\right) = \sigma^2$$

Uncorrelated:

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\sigma^2}{n}$$

Correlated:

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\sigma^2}{n} + \frac{n-1}{n}\rho\sigma^2$$

Assumptions

Ensemble members are more or less correct.

Ensemble members are diverse and (somewhat) independent.

How do we find these diverse models?

Ensemble learning – 2/4
How to combine predictions

Natural Computing

Gijs van Tulder

April 2020

Based on an earlier version by Jesse Krijthe.

How to combine the predictions?

Untrainable combiners
Mean, median, weighted mean, max, min, product.

Trainable combiners
Train a classifier in the intermediate space.

Combining predictions

M1	M2	M3	Mean	Max	Min	Prod
1	1	1	1	1 vs 0	1 vs 0	1 vs 0
1	1	0	2/3	1 vs 1	0 vs 0	0 vs 0
1	0	0	1/3	1 vs 1	0 vs 0	0 vs 0
0	0	0	0/3	0 vs 1	0 vs 1	0 vs 1

How to combine the predictions?

Untrainable combiners
Mean, median, weighted mean, max, min, product.

Trainable combiners
Train a classifier in the intermediate space.

Trainable combining / Stacking

Treat the predictions as features for a meta-classifier.

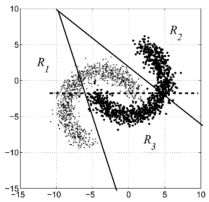
How to split the training data for the meta-classifier?

What about meta-meta-classifiers?

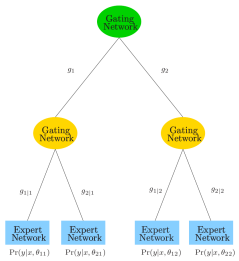
Sample	M1	M2	M3	True
1	1	1	1	1
2	1	1	0	1
3	1	0	0	0
4	0	0	0	0
5	0	1	0	1
6	0	1	1	0

Wolpert, "Stacked generalization", Neural networks, 1992.

Advanced combinations



Dynamic vs. global selection



Mixture of experts

Kuncheva (2004). Combining Pattern Classifers. Methods and Algorithms.
Hastie, Tibshirani and Friedman (2009). The Elements of Statistical Learning.

Ensemble learning – 3/4
How to create different models

Natural Computing

Gijs van Tulder

April 2020

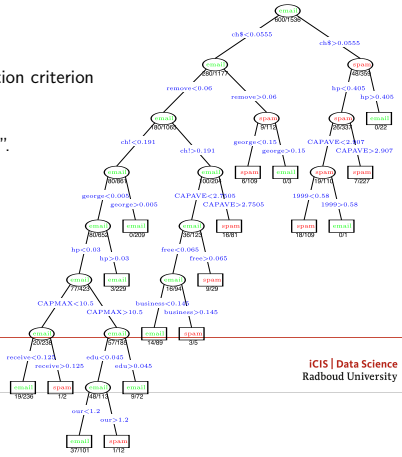
Based on an earlier version by Jesse Krijthe.

Decision tree

Greedy, minimising information criterion in each node.

Fast, flexible, "interpretable".

Difficult to control for overfitting.



Random forest

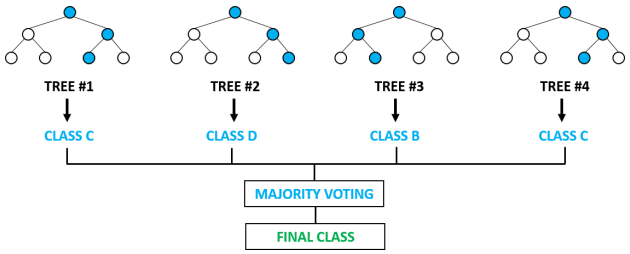


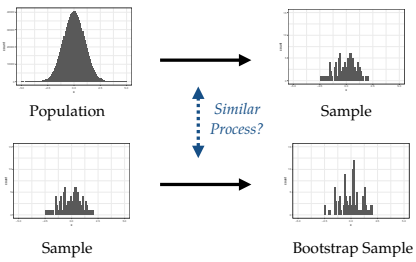
Image: globalsoftwaresupport.com

How do we get different trees?

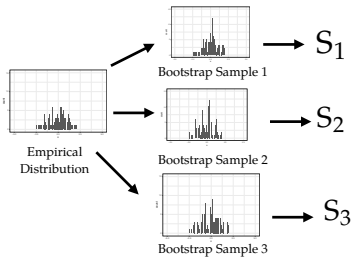
Using different subsets of training samples.

Using different subsets of features.

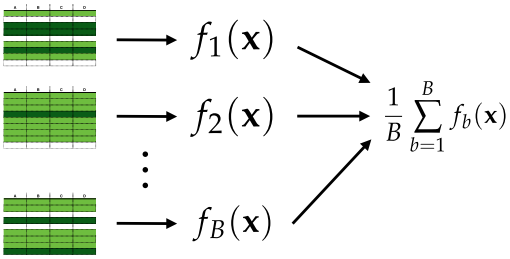
Bootstrapping



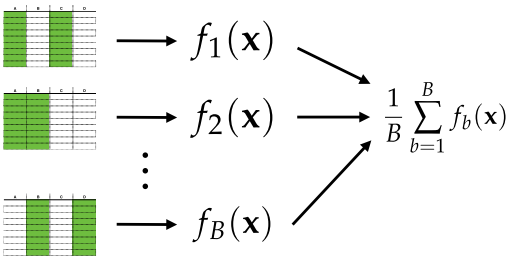
Bootstrapping



Bagging (bootstrap aggregating)



Random subspaces



Random forest

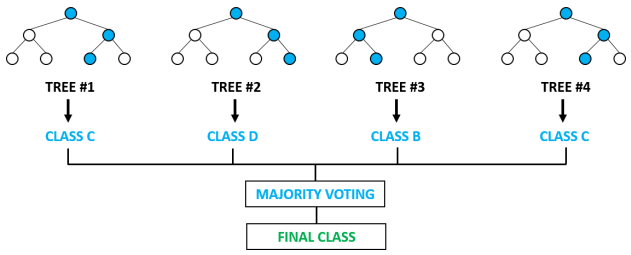


Image: globalsoftwaresupport.com

Parameters

Trees

Depth

Number of leaves

Minimum number of objects per node

Information criterion

Pruning

Forest

Number of trees

Size of subspaces considered

Pros and cons

Pros

Flexible/low bias

Works for many different types of data

Embarrassingly parallel

Produces out-of-bag (OOB) estimates

Scale invariant

Relatively few parameters

Cons

Harder to interpret than a single tree

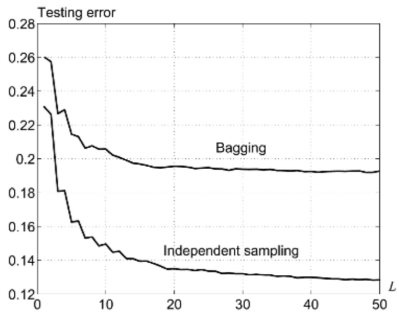
Computationally expensive

Bias-variance tradeoff

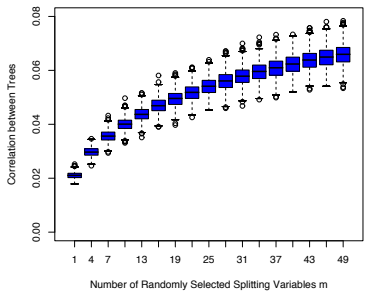
$$\underbrace{\mathbb{E}\left[(f-f^*)^2\right]}_{\text{Mean squared error}} = \underbrace{(\mathbb{E}[f]-f^*)^2}_{\text{Bias squared}} + \underbrace{\mathbb{E}\left[(f-\mathbb{E}[f]^*)^2\right]}_{\text{Variance}} + \underbrace{\sigma^2_{f^*}}_{\text{Irreducible error}}$$

From before:

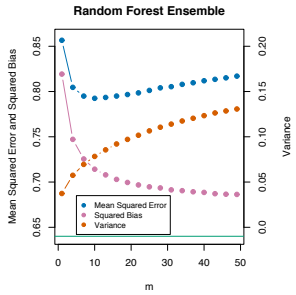
$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\sigma^2}{n} + \frac{n-1}{n}\rho\sigma^2$$



Kuncheva (2004). Combining Pattern Classifers. Methods and Algorithms.



Hastie, Tibshirani and Friedman (2009). The Elements of Statistical Learning.



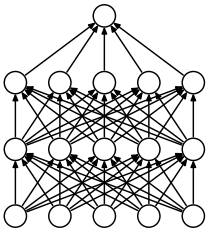
Hastie, Tibshirani and Friedman (2009). The Elements of Statistical Learning.

How to obtain different models?

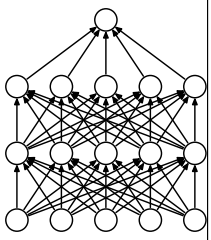
From a single model, we can generate a set of models by bootstrapping, random subspaces, or using the inherent stochasticity of the model.

Random forests use a combination of bootstrapping and random subspaces.

Dropout



(a) Standard Neural Net



(a) Standard Neural Net

Srivastava et al. (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". JMLR.

Ensemble learning – 4/4
Boosting

Natural Computing

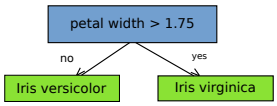
Gijs van Tulder

April 2020

Based on an earlier version by Jesse Krijthe.

Weak learners: Decision stumps

A weak learner is only slightly better than random guessing.

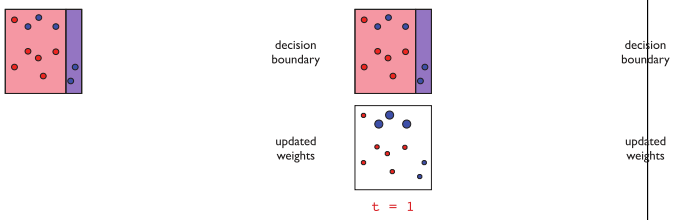


From weak to strong

Can we use a weak learner to construct a strong learner?

1. First apply a weak learner to our given dataset.
2. Train another weak learner that focusses on the mistakes of the first classifier (do this $M-1$ times).
3. Combine these in a smart way.

Adaptive reweighting and combining (arcing).



Mohri et al. (2012). Foundations of Machine Learning.

Algorithm 10.1 *AdaBoost.M1.*

1. Initialize the observation weights $w_i = 1/N,\ i = 1, 2, \ldots, N$.
2. For $m = 1$ to M:
   (a) Fit a classifier $G_m(x)$ to the training data using weights w_i.
   (b) Compute
   $$\mathrm{err}_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}.$$
   (c) Compute $\alpha_m = \log((1 - \mathrm{err}_m)/\mathrm{err}_m)$.
   (d) Set $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))],\ i = 1, 2, \ldots, N$.
3. Output $G(x) = \mathrm{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$.

Hastie, Tibshirani and Friedman (2009). The Elements of Statistical Learning.

Summary

Ensemble learning *combines decisions from different models* to obtain better predictions.

Bagging (bootstrap aggregating) trains models on different subsets of samples.

You can also use *random subspaces* of features to generate different models.

Boosting is an iterative method that trains models on earlier mistakes, by weighting the samples.

Ensemble learning

Natural Computing

Gijs van Tulder

April 2020

Based on an earlier version by Jesse Krijthe.

iCIS | Data Science
Radboud University