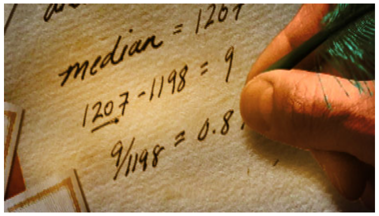
Ensemble learning

Chapter 11

The wisdom of the crowds

Sir Francis Galton discovered in the early 1900s that a collection of educated guesses can add up to very accurate predictions!



The paper in which he describes these findings:
Vox Populi. Nature 75, pages 450-451, 1907.
http://galton.org/essays/1900-1911/galton-1907-vox-populi.pdf

Image from http://www.pbs.org/wgbh/nova/physics/wisdom-crowds.html

Ensemble methods

Intuition: averaging measurements can lead to more reliable estimates

Need: an ensemble of different models from the same training data

How to achieve diversity?

Ensemble methods

Intuition: averaging measurements can lead to more reliable estimates

Need: an ensemble of different models from the same training data

How to achieve diversity?

Training models on random subsets or random subsets of features

Ensemble methods

The general strategy:

- Construct multiple, diverse predictive models from adapted versions of the training data
- Combine the predictions

5

Bootstrap samples

Bootstrap sample: sample with replacement from a dataset

The probability that a given example is not selected for a bootstrap sample of size n:

$$(1 - 1/n)^n$$

This has a limit as n goes to infinity: 1/e = 0.368

Conclusion: each bootstrap sample is likely to leave out about a third of the examples.

6

Bagging

Algorithm Bagging($D, T, \mathscr{A}$)

Input : data set D; ensemble size T; learning algorithm $\mathscr{A}$.
Output : ensemble of models whose predictions are to be combined by voting or averaging.

```
for t = 1 to T do
    build a bootstrap sample D_t from D by sampling |D| data points with replacement;
    run A on D_t to produce a model M_t;
end
return {M_t | 1 ≤ t ≤ T}
```

Breiman, Leo (1996). "Bagging predictors". Machine Learning 24 (2): 123–140.

7

Bagging

Comments:

How to combine the classifiers

- Average the raw classifier scores
- Convert to probabilities before averaging
- Voting

The decision boundary of a bagged classifier is more complex than that of the underlying classifiers.

8

Random forests

Random forests are a special case of bagging with the additional features:

- Use decision trees as the base classifier
- Sample the features for each bootstrapping sample

Random forests. L Breiman. Machine Learning, 2001.

Random forests

Algorithm RandomForest(D, T, d)

Input : data set D; ensemble size T; subspace dimension d.
Output : ensemble of tree models whose predictions are to be combined by voting or averaging.

```
for t = 1 to T do
    build a bootstrap sample D_t from D by sampling |D| data points with replacement;
    select d features at random and reduce dimensionality of D_t accordingly;
    train a tree model M_t on D_t without pruning;
end
return {M_t | 1 ≤ t ≤ T}
```


Random forests

Random forests are a special case of bagging with the additional features:

- Use decision trees as the base classifier
- Sample the features for each bootstrapping sample

- Variable importance: the values of the i^{th} feature are permuted among the training data and the out-of-bag error is computed on this perturbed data set. The importance score for is computed by averaging the difference in out-of-bag error before and after the permutation over all trees.
- Error estimation during training using out-of-bag data.

Why do committees work?

Let's focus on regression to illustrate this point.

The output produced by the committee:

$$f_{COM}(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} f_m(\mathbf{x})$$

Let's assume the output of each model can be represented as:

$$f_m(\mathbf{x}) = y(\mathbf{x}) + \epsilon_m(\mathbf{x})$$

The expected error of an individual model:

$$\mathbb{E}\left[(f_m(\mathbf{x}) - y(\mathbf{x}))^2\right] = \mathbb{E}\left[\epsilon_m(\mathbf{x})^2\right]$$

Why do committees work?

The average of the errors (corresponds to the situation where each model predicts indepdently):

$$E_{AV} = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\epsilon_m(\mathbf{x})^2\right]$$

Compare that to the expected error from the committee:

$$E_{COM} = \mathbb{E}\left[\left(\frac{1}{M}\sum_{m=1}^{M} f_m(\mathbf{x}) - y(\mathbf{x})\right)^2\right]$$
$$= \mathbb{E}\left[\left(\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\mathbf{x})\right)^2\right]$$

13

Why do committees work?

$$E_{AV} = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\epsilon_m(\mathbf{x})^2\right]$$
$$E_{COM} = \mathbb{E}\left[\left(\frac{1}{M}\sum_{m=1}^{M} f_m(\mathbf{x}) - y(\mathbf{x})\right)^2\right]$$
$$= \mathbb{E}\left[\left(\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\mathbf{x})\right)^2\right]$$

Let's assume:

$$\mathbb{E}\left[\epsilon_m(\mathbf{x})\right] = 0 \qquad \mathbb{E}\left[\epsilon_m(\mathbf{x})\epsilon_l(\mathbf{x})\right] = 0$$

Under this assumption: $E_{COM} = \frac{1}{M}E_{AV}$

14

Mixtures of experts

Bagging produces a result by averaging the predictions of the underlying classifiers.

Alternative: mixture of experts.

have each classifier be responsible for a small part of the input space

15

Boosting

Similar to bagging, but uses a more sophisticated method for constructing its diverse training sets.

Main ideas:

- Train the next classifier on examples that previous classifiers made errors on.
- Assign each classifier a confidence value that depends on its accuracy.

16

Boosting

Algorithm Boosting$(D, T, \mathscr{A})$

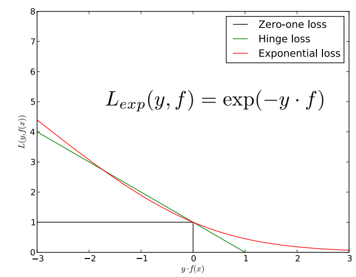
Input : data set D; ensemble size T; learning algorithm $\mathscr{A}$.
Output : weighted ensemble of models.

```
w_{1i} ← 1/|D| for all x_i ∈ D ;                                  // start with uniform weights
for t = 1 to T do
    run 𝒜 on D with weights w_{ti} to produce a model M_t;
    calculate weighted error ε_t;
    if ε_t ≥ 1/2 then
        set T ← t − 1 and break
    end
    α_t ← ½ ln((1−ε_t)/ε_t) ;                                       // confidence for this model
    w_{(t+1)i} ← w_{ti}/(2ε_t) for misclassified instances x_i ∈ D ;        // increase weight
    w_{(t+1)j} ← w_{tj}/(2(1−ε_t)) for correctly classified instances x_j ∈ D ;  // decrease
end
return M(x) = Σ_{t=1}^{T} α_t M_t(x)
```


Boosting and the exponential loss

What does boosting do?

Turns out that boosting minimizes the the exponential loss and leads to large margin solutions.



Robert E. Schapire, Yoav Freund, Peter Bartlett and Wee Sun Lee.
Boosting the margin: A new explanation for the effectiveness of voting methods.
The Annals of Statistics, 26(5):1651-1686, 1998.

Which classifier to boost?

Typically choose a simple classifier such as a "decision stump" or a simple linear classifier.

The bias-variance decomposition

The expected loss can be decomposed as follows (page 93):

expected loss = (bias)^2 + variance

bias – the extent to which the average prediction differs from the label

variance – the variability of the classifier when trained on different training sets

Very flexible models have very low bias, but high variance.

Rigid models have high bias and low variance.

The best model will achieve a good balance between the two.

Bagging vs boosting

expected loss = $(\text{bias})^2$ + variance

bias – the extent to which the average prediction is different differs from the label

variance – the variability of the classifier when trained on different training sets

Very flexible models have very low bias, but high variance.

Rigid models have high bias and low variance.

Bagging is a variance reduction technique, whereas boosting acts to reduce bias.

21