Resampling Methods and Ensemble Learners

Raia Hadsell
March 22, 2005
material from Duda, et al: ch. 9, and from slides by Yann LeCun

Overview

- Resampling Methods
  1. Jackknife
  2. Bootstrap
- Ensemble Learners
  1. Bagging
  2. Boosting
  3. Adaboost
  4. Querying

Resampling Methods

Resampling is used to estimate statistics: given a data set with an unknown probability distribution, how can the bias and variance be estimated?

How can the accuracy of other statistics be analyzed?

Resampling can also be used to improve classification (using ensemble learners).

Resampling Methods: mean and variance as usual

The usual mean of a data set S of n samples $x_{i..n}$ is

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The variance, or estimate of the accuracy of the mean, is

$$\hat{\sigma}^2 = \sum_{i=1}^{n} (x_i - \hat{\mu})^2.$$

Resampling Methods: motivation

- What if we want to know the median of the data set?

Easy.

- What if we want an estimate of the accuracy of the median (its spread)?

- How about the same estimate of accuracy for the mode, or the 37th percentile, or any other statistic?

- The jackknife and bootstrap methods are theoretically proven resampling techniques to produce accuracy estimates for arbitrary statistics, in the same way that the variance gives an estimate of accuracy for the mean.

Resampling Methods: Jackknife, or Leave-one-out

The leave-one-out mean is the average of the data set S if the ith sample is removed:

$$\mu_{(i)} = \frac{1}{n-1} \sum_{j \neq i}^{n} x_j.$$

and the Jackknife estimate of the mean is:

$$\mu_{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} \mu_{(i)},$$

i.e., the mean of the leave-one-out means.

Resampling Methods: Jackknife, or Leave-one-out

It is easy to prove that the Jackknife mean is the same as the traditional mean. Similarly, the Jackknife variance,

$$Var[\mu_{(\cdot)}] = \sum_{i=1}^{n} (\mu_{(i)} - \mu_{(\cdot)})^2$$

is equal to the traditional variance.

The advantage of using this formulation is that we can compute any statistic over S, then compute the statistic with each point left out. The spread of the statistic can then be estimated using the above equation, substituting the new statistic for $\mu_{(i)}$ and $\mu_{(\cdot)}$.

Resampling Methods: Bootstrap

A bootstrap data set B is created by randomly selecting n points from the training set S, with replacement. Since S only contains n points, the new bootstrap set almost always contains duplicate copies of samples in S. A number of such sets are produced and treated as independent sets.

The boostrap estimate of a statistic s, such as the mode or median, is the average of the statistic computed on each bootstrap set B:

$$\hat{s} = \frac{1}{b} \sum_{i=1}^{b} s(B_i).$$

Ensemble Learning Methods: Resampling for classifier training

- Resampling, along with other techniques, is useful in training classifiers, and generally such classifiers are called ensemble learners.
- The idea is to generate many learning machines that are slightly different from each other and make them vote.
- Democracy!
- The challenge: find a way to make each replica of the learning machine different from the others, yet useful.

Ensemble Learning Methods: Bagging

Bagging is the simplest ensemble method.

The name "bagging" comes from "bootstrap aggregation".

How to Bag:

- Create multiple bootstrap data sets, each one with $n' < n$ samples.
- Train a separate machine (a component classifier) on each data set.
- The classification is made by tallying the votes of the component classifiers.

Ensemble Learning Methods: Boosting

- The goal of Boosting is to "boost" the performance and accuracy of a learning algorithm.
- Successive component classifiers are trained with subsets of a data set. A new component's subset is created by selecting samples that were mis-classified by the existing components. Classification is by vote of the ensemble.
- Thus, new components are trained on samples that the previously trained component classifiers found difficult.
- This is a way to build multiple "experts" that specialize on different types of patterns.

Ensemble Learning Methods: 3 Component boosting

This is the boosting method given in the Duda, Hart, and Stork book:

1. Train the first component, G_1, on a training set S_1 (that is randomly drawn from the full training set S).
2. Select a second training set S_2 such that half the samples in S_2 are correctly classified by G_1 and half are incorrectly classified by G_1.
3. Train a second component classifier, G_2, on S_2.

Ensemble Learning Methods: 3 Component boosting

4. Create a third training set S_3 by taking only samples that G_1 and G_2 disagree on. Thus S_3 contains patterns not well represented by G_1 and G_2 together.
5. Train a third component classifier, G_3, on S_3.
6. Classification is based on a vote: if G_1 and G_2 agree, then use their decision. If G_1 and G_2 disagree, use the decision of G_3 instead.

Ensemble Learning Methods: Adaboost

- Adaboost ("adaptive boosting") is a very popular variation on basic boosting. Instead of using a set number of component learners, Adaboost allows you to keep adding component learners until some target training error is reached.
- Similarly to boosting, if a pattern is not accurately classified, then its chance of being picked for the next training subset is increased.
- In this way, Adaboost "focuses in" on informative or difficult patterns.

Ensemble Learning Methods: Adaboost

1. Given an n-sized training set S
2. Initialize a weight for each sample to $\frac{1}{n}$.
3. For each component G_i, do
   (a) Choose a training set S_i according to the weight on each sample.
   (b) Train a classifier G_i on S_i.
   (c) Compute E_i: the training error of G_i weighted by the sample weights.
   (d) Compute $\alpha_i =$ a measure of the accuracy of G_i.
   (e) Update each sample's weight- higher if G_i mis-classifies it.
4. Classification: take an α-weighted sum of the output of each component

Resampling Learning Methods: Learning with Queries

Learning with queries, a.k.a. active learning, a.k.a. interactive learning.

- Instead of the usual sample/label data set, consider a set of patterns that are unlabeled. Any pattern CAN be labeled, or queried, possibly for a cost, so the goal is to determine which patterns would be most informative to query.

- In Cost-based Learning, the queries have a pricetag, and the objective is to minimize an overall cost that depends both on classifier performance and total queries.

- Example: handwriting recognition

Learning with Queries: the process

•Start with a preliminary, weak classifier that's been trained with a small labeled set.

•Intuitively, the most informative samples to query will be close to the decision boundary. How to find them?

confidence-based query selection: the weak learner has discriminative functions for each category, and when a pattern is presented, it is deemed "informative" if the two largest functions give similar values.

voting-based query selection: Use this method in an ensemble machine. A new pattern is deemed "informative" if, when presented to several component classifiers, there is significant disagreement.

Learning with Queries

Artificial queries

If the number of unlabeled samples is small, sometimes samples can be generated that are close to the decision boundary (e.g., by rotating, skewing, scaling existing samples) or in uncharted areas of the decision space.