Ensemble Methods

A. Sanchez, F. Reverter and E. Vegas

Introduction to Ensembles

Some problems of *weak* learners.

- Decision trees have many good properties but some important drawbacks:
  - Smaller accuracy than competing alternatives
  - Very sensitive to small changes in data
  - Overall it makes the highly variable predictors
- Tree are not the only classifers to suffer from such problems.

Ensembles

- A common strategy to deal with these issues is to build repeated (weak learners) models on the same data and combine them to form a single result.
- These are called *ensemble* or consensus estimators/predictors.
- As a general rule, ensemble learners tend to improve the results obtained with the weak learners they are made of.

Ensemble methods

- Ensemble can be built on different learners but we will focus on those built on trees:
  - Bagging,
  - Random Forests,
  - Boosting,
  - Bayesian Trees.

Bagging: Aggregating predictors

Bagging: bootstrap aggregation

- Decision trees suffer from high variance when compared with other methods such as linear regression, especially when n/p is moderately large.
  - *NOTE: Write a small script to check this assertion*
- Given that high variance is intrinsec to the trees a possibility, suggested by Breimann (Breiman 1996), is to build multiple trees derived from the same dataset and, somehow, average them.

Averaging decreases variance

- Bagging relies, informally, on the idea that:
  - given $X \sim F()$, s.t. $Var_F(X) = \sigma^2$,
  - given a s.r.s. $X_1, ..., X_n$ from F then
  - if $\overline{X} = \frac{1}{N}\sum_{i=1}^{n} X_i$ then $var_F(\overline{X}) = \sigma^2/n$.
- That is, *relying on the sample mean instead of on simple observations decreases variance by a factor of* n.

Averaging trees …

Two questions arise here:

1. How to go from X to $X_1, ..., X_n$?

- This will be done using *bootstrap resampling*.

2. What means "averaging" in this context.

- Depending on the type of tree:
  - Average predictions for regression trees.
  - Majority voting for classification trees.

The bootstrap

- *Bootstrap* methods were introduced by Bradley Efron in 1979 (Efron 1979) to estimate the standard error of a statistic.
- The success of the idea lied in that the procedure was presented as "automatic' ', that is:
  - instead of having to do complex calculations,
  - it allowed to approximate them using computer simulation.
- Some people called it "the end of mathematical statistics' '.

Bootstrap Applications

- The bootstrap has been applied to almost any problem in Statistics.
  - Computing standard errors,
  - Bias estimation and adjustment,
  - Confidence intervals,
  - Significance tests, …
- We begin with the easiest and best known case: *estimating the standard error (that is the square root of the variance) of an estimator*.

Precision of an estimate (1)

- Assume we want to estimate some parameter θ, that can be expressed as $\theta(F)$, where F is the distribution function of each X_i in $(X_1, X_2, ..., X_n)$.
- For example:

$$
\begin{aligned}
\theta &= E_F(X) = \theta(F) \\
\theta &= Med(X) = \{m : P_F(X \leq m) = 1/2\} = \theta(F).
\end{aligned}
$$

Plug-in estimates

- To estimate $\theta(F)$ we usually rely on *plug-in estimators*: $\hat{\theta} = \theta(F_n)$:

$$
\begin{aligned}
\hat{\theta} &= \overline{X} = \int X dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} x_i = \theta(F_n) \\
\hat{\theta} &= \widehat{Med}(X) = \{m : \frac{\# x_i \leq m}{n} = 1/2\} = \theta(F_n)
\end{aligned}
$$