Principles of Statistical Data Mining 11/13/2018

Lecture 1: Ensemble Learning:Bagging

Professor: Dr Ernest Fokoué Scribe: Miss Sailee Rumao

I. Introduction

When different models which are qualitatively very different and have almost similar model selection scores (eg. BIC), **Model selection** is not necessarily the most ideal way for classification because when you choose one , you are making strong rejection against other variables that you do not choose.
This is one of the motivation for **Model Aggregation**.
We also know that,

$$V(\bar{\mathrm{x}}) < V(\mathrm{x}_i) \qquad (\forall\, i = 1, 2, ...n) \tag{1}$$

This is another motivation for Model Aggregation. In such a case , ensemble (Aggregated Model) will have very low variance which results in low prediction error based on the **Bias-Variance Decomposition**.

II. Ensemble Learning

It is a method used for increasing model accuracy by aggregating k learned models $C_1, C_2, \cdots, C_k$ (classifiers or predictors) with an aim of creating an improved composite model C^*.
Here, C : Classifier.

Consider $\mathbb{D} = \left\{ Z_1, Z_2, \cdots, Z_n \right\}$ where $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ where are two random variables.
Assume f : function that models relationship between $\mathcal{X}$ and $\mathcal{Y}$
$\hat{f_1}(\cdot), \hat{f_2}(\cdot), \cdots, \hat{f_m}(\cdot)$: M different estimators of f
α_m : Weights that measure relative importance of each estimator $\hat{f_m}(\cdot)$ m=1,2,3,..M
Our goal is to find an estimator of f with the smallest prediction error.

Mathematically,

$$\widehat{f}^{(agg)}(\cdot) = \sum_{m=1}^{M} \alpha_m \hat{f_m}(\cdot) \quad i = 1, 2, \cdots M \tag{2}$$

Each of the estimator $\hat{f_m}(.)$ is called **Base Learner.**
Note: If the base estimators $\hat{f_m}(.)$ are strongly correlated then the performance of the ensemble is likely to decrease.

The two techniques/Methods of Ensemble learning are:
1. Bagging
2. Boosting

II.1 Bagging(Leo Breiman (1996))

Bagging is also known as Bootstrap aggregation.

The steps for bagging are as follows:

1. Drawing $\mathbb{D}^{(m)} = (\mathbf{x}_i^{(m)}, \mathbf{y}_i^{(m)})$ with replacement from D.
This is called bootstrap sampling.

2. Build base learners $\hat{f_m}(\cdot)$ using $\mathbb{D}^{(m)}$ and the chosen learning machine.

3. Then if all base learners have equal weights i.e $\hat{\alpha_m} = \frac{1}{M}$ (in case of bagging),

$$\hat{f}^{(agg)}(\cdot) = \frac{1}{M}\sum_{m=1}^{M} \hat{f_m}(\cdot) \tag{3}$$

Note:
1. Many $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)$ from $\mathbb{D}$ are replaced many times in $\mathbb{D}^{(m)}$

2. $|\mathbb{D}^{(m)}| = n$ (which is the size of $\mathbb{D}$)
3. In Regression Model avergaing is obvious and is given as,

$$\hat{f}^{(agg)}(X_{new}) \overset{Reg}{=} \frac{1}{M}\sum_{m=1}^{M} \hat{f}^{(m)}(X_{new}) \tag{4}$$

In classification we use the majority rule (for the class labels obtained) to make a decision which is given as follows:

$$\hat{f}^{(agg)}(X_{new}) \overset{Classif}{=} \underset{g=1,2,3...G}{\texttt{argmax}} \left\{ \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}(\hat{f}^{(m)}(X_{new}) = g) \right\} \tag{5}$$

4. Bias of Machine depends on the bias of the base learners.

Out of Bag Error

Consider $\mathbf{z} \in \mathbb{D}$ be a random pair in $\mathbb{D}$ (original dataset of size n) and
Consider bootstraped sample $\mathbb{D}^{(m)}$ of size n , then,

$Pr(\mathbf{z} \in \mathbb{D}^{(m)})$ = Proportion of observations from $\mathbb{D}$ present in $\mathbb{D}^{(m)} = 1 - (1 - \frac{1}{n})^n$ .....(*)
$Pr(\mathbf{z} \notin \mathbb{D}^{(m)})$ = Proportion of observations from $\mathbb{D}$ present in $\mathbb{D}^{(m)} = (1 - \frac{1}{n})^n = Pr[O_n]$ .....(**)

As $n \to \infty \quad , Pr[O_n] \to e^{-}1 = 0.37$

This implies that approximately one third of the training set is not used to build mth bootstraped base learner. This proportion is used to calculate the out of bag error.(used in random forest to estimate variable importance.)

Now let $\hat{f}^{(m)}(.)$ be the base learner from $\mathbb{D}^{(m)}$.
Then the out of bag error of $\hat{f}^{(m)}(\cdot)$ is ,

$$err_{OOB}(\hat{f}^{(m)}(\cdot)) = \frac{\sum_{i=1}^{n} \mathbb{1}(\mathbf{z}_i \notin \mathbb{D}^{(m)})\, \ell(\mathbf{y}_i, \hat{f}^{(m)}(\mathbf{x}_i))}{\sum_{i=1}^{n} \mathbb{1}(\mathbf{z}_i \notin \mathbb{D}^{(m)})} \tag{6}$$

where,
$\mathbb{1}(\cdot)$: Indicator function
$\ell(\cdot, \cdot)$:Loss function
Thus, the OOB error for the ensemble is,

$$err_{OOB}(\hat{f}^{(agg)}(\cdot)) = \frac{1}{M} \sum_{m=1}^{M} err_{OOB}(\hat{f}^{(m)}(\cdot)) \tag{7}$$

Note:
1. Out of bag error is a good estimate of test error.
2. $\lceil e^{-}1 \rceil$ = size of the out of bag sample(OOB sample) $\approx$ Number of observations from $\mathbb{D}$ not in $\mathbb{D}^{(m)}$

Properties of Bagged Ensembes:

1. If $\hat{f}^{(m)}(.)$ is an unbiased estimator of f^* i.e $\mathbb{E}[\hat{f}^{(m)}(\cdot)] = f^*$ then
$\hat{f}^{(agg)}(\cdot)$ will also be unbiased i.e $\mathbb{E}[\hat{f}^{(agg)}(\cdot)] = f^*$

Expectation of final predictor is,

$$\mathbb{E}[\hat{f}^{(agg)}(.)] \overset{bagging}{=} \mathbb{E}[\frac{1}{M}\sum_{m=1}^{M}\hat{f}^{(m)}(\cdot)] = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}[\hat{f}^{(m)}(\cdot)] \tag{8}$$

Thus the above statement.

2. Assume,

$\mathbb{V}(\hat{f}^{(m)}(\cdot)) = \sigma^2 \qquad \forall\, m = (1, 2 \cdots, M)$
$\text{Cov}(\hat{f}^{(l)}(\cdot), \hat{f}^{(m)}(\cdot)) = \tau$
Then,

$$\mathbb{V}(\hat{f}^{(agg)}(\cdot)) = \mathbb{V}(\frac{1}{M}\sum_{m=1}^{M}\hat{f}^{(m)}(\cdot)) = \frac{1}{M^2}\mathbb{V}(\sum_{m=1}^{M}\hat{f}^{(m)}(\cdot)) + \frac{1}{M^2}\sum_{l=1}^{M}\sum_{m=1}^{M}Cov(\hat{f}^{(l)}(.), \hat{f}^{(m)}(\cdot)) \tag{9}$$

Now, if $\tau = 0$
then ,

$$\mathbb{V}(\hat{f}^{(agg)}(.)) = \frac{1}{M^2}\mathbb{V}(\sum_{m=1}^{M}\hat{f}^{(m)}(\cdot)) = \frac{\sigma^2}{M} \tag{10}$$

and $\lim_{M\to\infty}(\mathbb{V}(\hat{f}^{(agg)(.)})) = 0$
i.e As M tends to infinity , variance of the aggregate Machine decreases without any decrease in its bias resulting in bias-varinace tradeoff and therefore leading to an optimal predictive performance.
Inother words, if the individual machines are unbiased and they are near orthogonal , the aggregate machine will become a consistent estimator of the true function.

Note: Bagging always uses the whole p-dimensional vector X_i. This can be bad for n««p (curse of dimensionality) and for variable selection.

Random Forest

Random Forest overcomes the drawback of Curse of Dimensionality in Bagging.

Bagging is performed in Random forest in the following steps:
1.

1. Pick a subset of d variables from the original p-variables
   $(\mathbf{x}_{j1}, \mathbf{x}_{j2} \cdots, \mathbf{x}_{jd})$: Subset of variables (where d $<$ p)
   Here $\mathbf{x}_j k$ are chosen without replacement
   $\mathbb{D}^{(m)}$ is an nxd matrix
2. Choose a bootstrap sample
3. Construct base learner

2. Device for Variable Importance

   1. Take out of bag sample of size S = $\lceil e^{-1} n \rceil$
   2. After $\hat{f}(m)(\cdot)$ is built ,calculate out of bag error i.e $\hat{\mathbb{E}}_{OOB}$ (raw)
   3. For k=1 to d,
      (a) Perform permutation of column j_k in OOB sample (shuffle the entries in column j_k)
      (b) Compute $\hat{\mathbb{E}}^{(k)}_{OOB}$
      (c) Compute $\hat{\mathbb{E}}_{OOB} - \hat{\mathbb{E}}^{(k)}_{OOB}$

This Mechanism allows Random forest to measure variable importance.

Advantages of Random Forest:

1. Achieves a more powerful and flexible machine.
2. Gives a measure of variable importance.

II.2 Boosting

The imporatnce of boosting lies in combining various base learners assumed to be weak(descent machines) in the sense that each of them is just slightly better than random guessing.
Boosting uses complicated α.

We'll look at the working of Adaptive boosting which is a type of Boosting in the next lecture