Lecture 10 Decision Trees and Ensemble Methods

1. Decision Trees

Greedy, Top-Down, Recursive

Notation: Region R_p;

Looking for a split S_p.

$$S_p(j,t) = (\{x \mid x_j < t,\ x \in R_p\},\ \{x \mid x_j \geq t,\ x \in R_p\})$$

($\{x \mid x_j < t,\ x \in R_p\} \to R_1$, $\{x \mid x_j \geq t,\ x \in R_p\} \to R_2$)

1) How to choose Split?

Define $L(R)$: loss on R

Given C classes, define $\hat{p}_c$ to be the proportion of examples in R that are of class C.

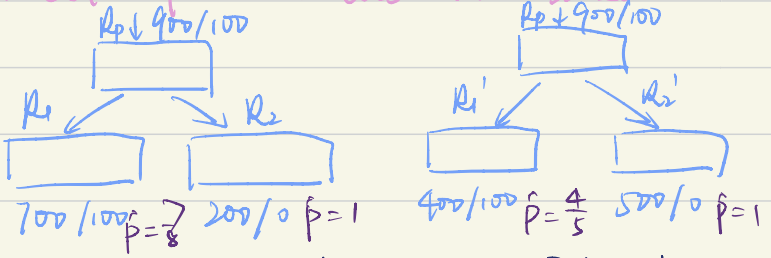
$L_{misclass} = 1 - \max_c \hat{p}_c$

parent loss　　children loss

$$\max_{j,t} L(R_p) - (L(R_1) + L(R_2))$$

But, misclassification loss has issues:

e.g.



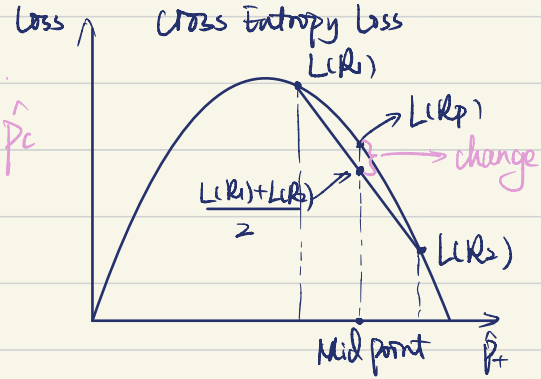
$L(R_1) + L(R_2) = 100 + 0 = 100$

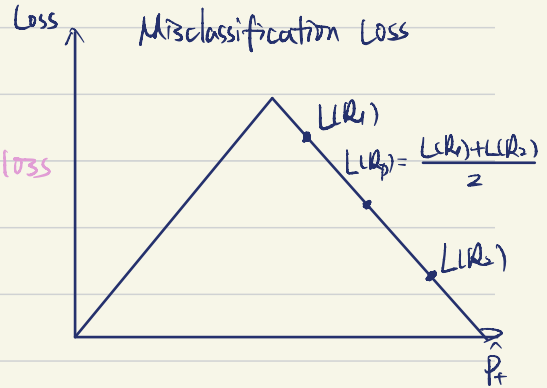
$L(R_1') + L(R_2') = 100 + 0 = 100$

$L(R_p) = 100$

- Cross Entropy Loss.

$$L_{cross} = -\sum_c \hat{p}_c \log_2 \hat{p}_c$$





2. Regression Trees

Predict: $\hat{y}_m = \frac{\sum_{i \in R_m} y_i}{|R_m|}$

$L_{squared} = \frac{\sum_{i \in R_m} (y_i - \hat{y}_m)^2}{|R_m|}$

3. Categorical Variables

n categories $\Rightarrow 2^n$ possible splits

4. Regularization of Decision Trees

1) minimize leaf size
2) maximize depth
3) maximize # nodes
4) minimize decrease in loss

5) Pruning (misclassification with validation sets)

5. Runtime

	Test time	Train Time
n examples	$O(d)$	Each point is part of $O(d)$ nodes
f features	$d < \log_2 n$	Cost of point at each node is $O(f)$
d depth		total cost is $O(nfd)$
		Data matrix is of size nf

6. No additive structure

7. Adv & Disadv.

Adv

Disadv

① Easy to explain
② Interpretable
③ Categorical variables
④ Fast

① High Variance
② Bad at additive
③ Low predictive accuracy

8. Ensembling

1) Take X_i's which are random variables (RV) that are independently identically distributed (iid)

$Var(X_i) = \sigma^2$ $\quad Var(\bar{X}) = Var(\frac{1}{n}\sum_i X_i) = \frac{1}{n}\sigma^2$

2) Drop independence assumption $\Rightarrow X_i$ is id.

X_i's correlated by ρ

$Var(\bar{X}) = \rho\sigma^2 + \frac{1-\rho}{n}\sigma^2$

- Ways to Ensemble

1) different algorithms
2) different training sets
3) Bagging (Random Forest)
4) Boosting (Adaboost, xgboost)

9. Bagging – Bootstrap Aggregation

Have a true population P

Training Set $S \sim P$

Assume $P = S$

Bootstrap samples $Z \sim S$

Bootstrap samples $Z_1, \cdots Z_m$

Train Model G_m on Z_m, $G(x) = \frac{\sum_{m=1}^{M} G_m(x)}{M}$

10. Bias - Variance Analysis

$Var(\bar{x}) = \rho\sigma^2 + \frac{1-\rho}{M}\sigma^2$

Bootstrapping is driving down ρ.

More M → less variance

Bias slightly increases because of random sampling

* Decision Trees have high variance, low bias. ⇒ ideal for bagging

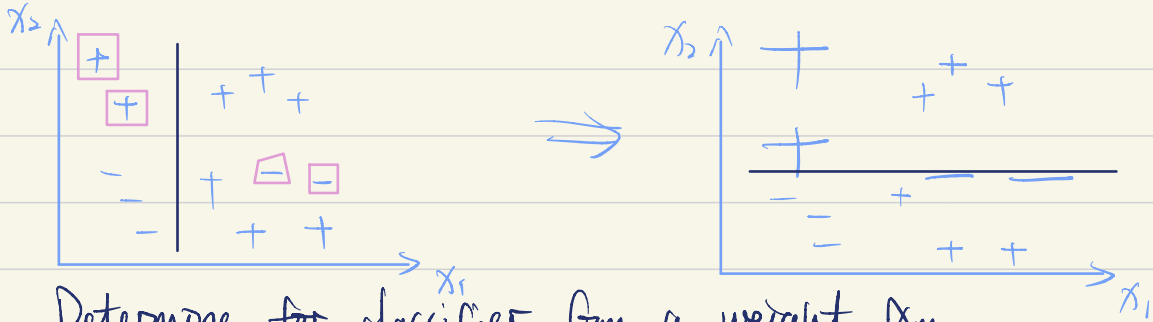
11. Random Forest

At each split, consider only a fraction of your total feature.

decrease ρ and decorrelate models

12. Boosting (decrease bias; additive)

example:



Determine for classifier G_m a weight α_m proportional $\log\left(\frac{1-err_m}{err_m}\right)$

- Adaboost : $G(x) = \sum_m \alpha_m G_m$

Each G_m trained on reweighted training set