MATH 390.4 / 650.2 Spring 2020 Homework #5

Tziporah Horowitz

Monday 11th May, 2020

1 The CART Algorithms

(a) [easy] Write down the step-by-step $\mathcal{A}$ for regression trees.

Let the dataset be all the data,

(a) Consider every possible orthogonal-to-axis split ($X_j \leq X_{ij}\ \forall\ j = 1, \ldots, p$, $i \in \{1, \ldots, n-1\}$). Compute SSE_l and SSE_r (the SEE's in the left node and the right node). Select the node where $SSE_{weighted} = \frac{n_l SSE_l + n_r SSE_r}{n_l + n_r}$ is the smallest, i.e. create an inner node with that split rule, a left leaf with $\hat{\boldsymbol{y}} = \bar{y}_l$, and a right leaf with $\hat{\boldsymbol{y}} = \bar{y}_r$.

(b) If $n_l > N_0$ (a hyperparameter), set the dataset on the left of the partition and run (a) on it. If $n_r > N_0$, set the dataset on the right of the partition and run (a) on it.

(b) [difficult] Describe $\mathcal{H}$ for regression trees. This is very difficult but doable. If you can't get it in mathematical form, describe it as best as you can in English.

$\mathcal{H}$ is the set of all split rules.

(c) [harder] Think of another "leaf assignment" rule besides the average of the responses in the node that makes sense.

$$\hat{\boldsymbol{y}} = \text{Median}[\boldsymbol{y}]$$

(d) [harder] Assume the y values are unique in $\mathbb{D}$. Imagine if $N_0 = 1$ so that each leaf gets one observation and its $\hat{\boldsymbol{y}} = y_i$ (where i denotes the number of the observation that lands in the leaf) and thus it's very overfit and needs to be "regularized". Write up an algorithm that finds the optimal tree by pruning one node at a time iteratively. "Prune" means to identify an inner node whose daughter nodes are both leaves and deleting both daughter nodes and converting the inner node into a leaf whose $\hat{\boldsymbol{y}}$ becomes the average of the responses in the observations that were in the deleted daughter nodes.

This is an example of a "backwards stepwise procedure" i.e. the iterations transition from more complex to less complex models.

(e) [difficult] Provide an example of an $f(\boldsymbol{x})$ relationship with medium noise δ where vanilla OLS would beat regression trees in oos predictive accuracy. Hint: this is a trick question.

If $f(\boldsymbol{x})$ is a linear equation, OLS will be able to model it with low error.

$$f(\boldsymbol{x}) = b_0 + b_1 x_1 + \ldots + b_p x_p$$

(f) [easy] Write down the step-by-step $\mathcal{A}$ for classification trees. Feel free to reference steps in (a).

(a) Consider every possible orthogonal-to-axis split ($X_j \leq X_{ij}\ \forall\ j = 1, \ldots, p$, $i \in \{1, \ldots, n-1\}$). Rather than selecting the node where $SSE_{weighted}$ is the smallest, select the node where $Gini_N = \frac{n_l Gini_l + n_r Gini_r}{n_l + n_r}$ is the smallest, where $Gini_l = \sum_{k=1}^{k} \hat{p}_k(1 - \hat{p}_k)$ and $\hat{p}_k = \sum_{i=1}^{n_l} \mathbb{1}_{y_i = c_k}$. Leaf assignment will be based on the mode instead of the mean.

(b) Same as step (b) in question (a).

(g) [difficult] Think of another objective function that makes sense besides the Gini that can be used to compare the "quality" of splits within inner nodes of a classification tree.

2 Probability Estimation and Asymmetric Cost Modeling

(a) [easy] Why is logistic regression an example of a "generalized linear model" (glm)?

Because we retain $\boldsymbol{w} \cdot \boldsymbol{x}$ in the hypothesis set.

(b) [easy] What is $\mathcal{H}_{pr}$ for the probability estimation algorithm that employs the linear model in the covariates with logistic link function?

$$\mathcal{H}_{pr} = \left\{ \frac{1}{1 + e^{-\boldsymbol{w} \cdot \boldsymbol{x}}} : \boldsymbol{w} \in \mathbb{R}^{p+1} \right\}$$

(c) [easy] If logistic regression predicts 3.1415 for a new $\boldsymbol{x}_*$, what is the probability estimate that $y = 1$ for this $\boldsymbol{x}_*$?

$$p \approx 1$$

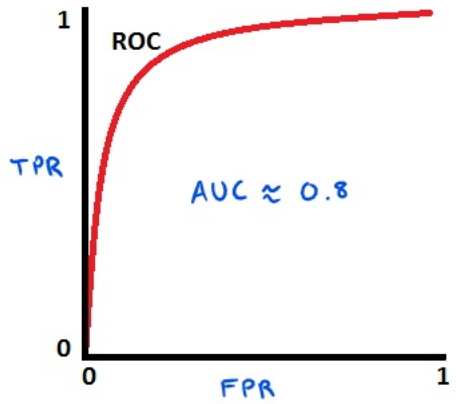
(d) [harder] What is $\mathcal{H}_{pr}$ for the probability estimation algorithm that employs the linear model in the covariates with cloglog link function?

$$\mathcal{H}_{pr} = \left\{ 1 - e^{-e^{\boldsymbol{w} \cdot \boldsymbol{x}}} : \boldsymbol{w} \in \mathbb{R}^{p+1} \right\}$$

(e) [difficult] Generalize linear probability estimation to the case where $\mathcal{Y} = \{C_1, C_2, C_3\}$. Use the logistic link function like in logistic regression. Write down the objective function that you would numerically maximize. This objective function is one that is argmax'd over the parameters (you define what these parameters are — that is part of the question).

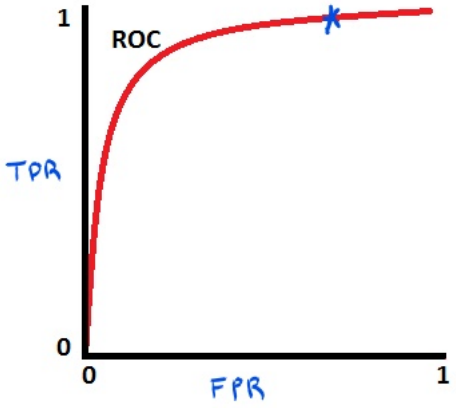
Once you get the answer you can see how this easily goes to $K > 3$ response categories. The algorithm for general K is known as "multinomial logistic regression", "polytomous LR", "multiclass LR", "softmax regression", "multinomial logit" (mlogit), the "maximum entropy" (MaxEnt) classifier, and the "conditional maximum entropy model". You can inflate your resume with lots of jazz by doing this one question!

(f) [easy] Graph a canonical ROC and label the axes. In your drawing estimate AUC. Explain very clearly what is measured by the x axis and the y axis.



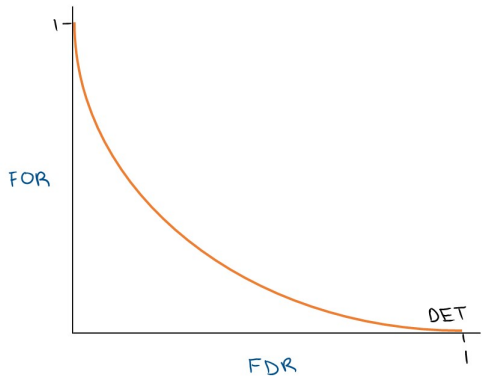
FPR is the rate at which negatives are discovered as false positives. TPR is the rate at which positives are discovered as true positives.

(g) [easy] Pick one point on your ROC curve from the previous question. Explain a situation why you would employ this model.



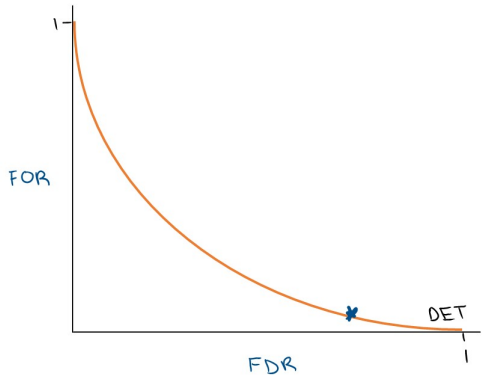
We care more about the reward of predicting true negatives than the cost of a false positives.

(h) [easy] Graph a canonical DET curve and label the axes. Explain very clearly what is measured by the x axis and the y axis.



FDR is the proportion of discovered positives that are false. FOR is the proportion of discovered negatives that are false.

(i) [easy] Pick one point on your DET curve from the previous question. Explain a situation why you would employ this model.



We care more about the cost of a false negative than the cost of a false positive.

(j) [difficult] The line of random guessing on the ROC curve is the diagonal line with slope one extending from the origin. What is the corresponding line of random guessing in the DET curve? This is not easy...

The line of random guessing on the DET curve would be the diagonal line with the y intercept of 1 and the slope -1.

3 Bias-Variance Decomposition

Assume the two assumptions from the notes about the random variable model that produces the δ values, the error due to ignorance.

(a) [easy] Write down (do not derive) the decomposition of MSE for a given $\boldsymbol{x}_*$ where $\mathbb{D}$ is assumed fixed but the response associated with $\boldsymbol{x}_*$ is assumed random.

$$MSE(\boldsymbol{x}_*) = \sigma^2 + \big(f(\boldsymbol{x}_*) - g(\boldsymbol{x}_*)\big)^2$$

(b) [easy] Write down (do not derive) the decomposition of MSE for a given $\boldsymbol{x}_*$ where the responses in $\mathbb{D}$ is random but the $\boldsymbol{X}$ matrix is assumed fixed and the response associated with $\boldsymbol{x}_*$ is assumed random like previously.

$$MSE(\boldsymbol{x}_*) = \sigma^2 + \text{Bias}[G]^2 + \mathbb{V}\text{ar}\,[G]$$

(c) [easy] Write down (do not derive) the decomposition of MSE for general predictions of a phenomenon where all quantities are considered random.

$$MSE = \sigma^2 + \mathbb{E}\left[\text{Bias}[G]^2\right] + \mathbb{E}\left[\mathbb{V}\text{ar}\,[G]\right]$$

(d) [difficult] Why is it in (a) there is only a "bias" but no "variance" term? Why did the additional source of randomness in (b) spawn the variance term, a new source of error?

There is no variance term in (a) because there is only one model. In (b), the randomness in $\mathbb{D}$ creates different δ's so there is variance in the distribution of multiple models.

(e) [harder] A high bias / low variance algorithm is underfit or overfit?

Underfit.

(f) [harder] A low bias / high variance algorithm is underfit or overfit?

Overfit.

(g) [harder] Explain why bagging reduces MSE for "free" regardless of the algorithm employed.

It allows $\mathcal{H}$ to be very expressive, so the Bias$[g_i]$ is approximately 0 for all i.

(h) [harder] Explain why RF reduces MSE atop bagging M trees and specifically mention the target that it attacks in the MSE decomposition formula and why it's able to reduce that target.

Since bagging reduces the bias to approximately 0,

$$MSE \approx \sigma^2 + \rho \mathbb{E}\left[\mathbb{V}\text{ar}\left[g_{(m)}\right]\right]$$

where ρ is the average correlation between two trees. Random Forest attacks ρ by splitting on a subset of features of size $p_{try} < p$ after node construction, thereby making the trees more different from one another.

(i) [difficult] When can RF lose to bagging M trees? Hint: setting this critical hyperparameter too low will do the trick.

If the hyperparameter is too low, the bias will increase too much.

4 Lasso, Ridge, and the Elastic Net

(a) [easy] Write down the objective function to be minimized for ridge. Use λ as the hyperparameter.

$$\boldsymbol{b} = \underset{\boldsymbol{w} \in \mathbb{R}^{p+1}}{\arg\min} \left\{ SSE + \lambda ||\boldsymbol{w}||^2 \right\}$$

(b) [easy] Write down the objective function to be minimized for lasso. Use λ as the hyperparameter.

$$\boldsymbol{b} = \underset{\boldsymbol{w} \in \mathbb{R}^{p+1}}{\arg\min} \left\{ SSE + \lambda ||\boldsymbol{w}|| \right\}$$

(c) [easy] We spoke in class about when ridge and lasso are employed. Based on this discussion, why should we restrict $\lambda > 0$?

If $\lambda = 0$, $\boldsymbol{X}^\top \boldsymbol{X}$ is not invertable. If $\lambda < 0$, $\boldsymbol{b}$ can be 0.

(d) [harder] Why is lasso sometimes used a preprocessing step to remove variables that likely are not important in predicting the response?

Lasso sets many elements of $\boldsymbol{b}$ to 0.

(e) [easy] Assume $\boldsymbol{X}$ is orthonormal. One can derive lasso in closed form. Copy the answer from the wikipedia page. Compare lasso to OLS.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \cdot \max\left(0, 1 - \frac{N\lambda}{|(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}|}\right)$$

(f) [harder] Write down the objective function to be minimized for the elastic net. Use α and λ as the hyperparameters.

$$\boldsymbol{b} = \underset{\boldsymbol{w} \in \mathbb{R}^{p+1}}{\arg\min} \left\{ SSE + \lambda\big(\alpha||\boldsymbol{w}||_1 + (1-\alpha)||\boldsymbol{w}||_2^2\big) \right\}$$

(g) [easy] We spoke in class about the concept of the elastic net. Based on this discussion, why should we restrict $\alpha \in (0, 1)$?

So that both ridge and lasso will be factored into the regression.

5 Missingness

(a) [easy] What are the three missing data mechanisms? Provide an example when each occurs (i.e., a real world situation).

- *Missing Completely at Random* (MCAR). Ex: there are random missing values in the data that have nothing to do with one another.
- *Missing at Random* (MAR). Ex: Men are more likely to report their weight than women.
- *Not Missing at Random* (NMAR). Ex: retired seniors not filling out hourly wage in a survey.

(b) [easy] Why is listwise-deletion a terrible idea to employ in your $\mathbb{D}$ when doing supervised learning?

You can end up dropping too much important data.

(c) [easy] Why is it good practice to augment $\mathbb{D}$ to include missingness dummies? In other words, why would this increase oos predictive accuracy?

The dummy variables take the weight of the missing data.

(d) [easy] To impute missing values in $\mathbb{D}$, what is a good default strategy and why?

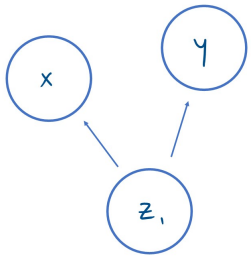
Use missForest because it guarantees convergence of the imputed values.

6 Correlation-Causation and Interpretation of OLS coefficients.

(a) [easy] Consider a fitted OLS model for y with features x_1, x_2, ..., x_p. Provide the most correct interpretation of the quantity b_1 you can.

When computing two "mutually observed" independent observations A and B, sampled in the same fashion as observations in $\mathbb{D}$, when A has an x_1 measurement one unit larger than B's x_1 measurement but both A and B share the same features $x_1, \ldots, x_p$ then A is predicted to have a response y that differs on average by b_1 units from the response y of B.

(b) [easy] If x and y are correlated but their relationship isn't causal, draw a diagram below that includes z.



(c) [easy] To show that x is causal for y, what specifically has to be demonstrated? Answer with a couple of sentences.

To have a causal relationship, x must be connected to y, whether directly or indirectly. Often in modeling, the causal relationship is indirect, i.e. x causes z and z causes y along with other factors.

(d) [harder] If we fit a model for y using x_1, x_2, ..., x_7, provide an example real-world illustration of the causal diagram for y including the z_1, z_2, z_3.