MATH 342W / 642 / RM 742 Spring 2024 HW #5

Osman Khan

Wednesday 15th May, 2024

Problem 1

These are some questions related to probability estimation modeling and asymmetric cost modeling.

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

Logistic regression is an example of a generalized linear model as it retains $\vec{w} \cdot \vec{x}$ but manipulates it in some way.

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

$\mathcal{H}_{pr} = \{\phi(u = \vec{w} \cdot \vec{x}) = \frac{e^u}{1+e^u} : (\vec{w} \cdot \vec{x}) \in \mathbb{R}^{p+1}\}$

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

$\phi(3.1415) = \frac{e^{3.1415}}{1+e^{3.1415}} \approx 0.96$

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

From Wikipedia, cloglog $= \log(-\log(1-p))$, so we find inverse:
$u = \log(-\log(1-p))$
$= 10^u = -\log(1-p)$
$= 10^u = \log(1-p)^{-1}$
$= 10^{10^u} = \frac{1}{1-p}$
$= 1 - p = \frac{1}{10^{10^u}}$
$= p = 1 - \frac{1}{10^{10^u}}$
$\therefore \mathcal{H}_{pr} = \phi(u) = 1 - \frac{1}{10^{10^u}}$
If Log is actually Ln, then we have: $\mathcal{H}_{pr} = \phi(u) = 1 - \frac{1}{e^{e^u}} = 1 - e^{-e^u}$

(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

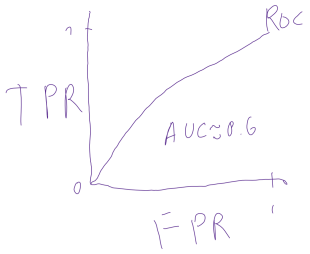

Figure 1: 1.f

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!

$\phi(i) = argmax \frac{e^i}{\sum_{j=1}^{3} e^j}$

(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.

The x axis measures the negatives that are false negatives while the y axis measures the positives that are true positives.

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

At x=0.2, profit of predicting true negatives is less than the cost of false positive.

(h) [harder] Graph a canonical DET curve and label the axes. Explain very clearly what is measured by the x axis and the y axis. Make sure the DET curve's intersections with the axes is correct.

The x axis measures the proportion of predicted positives that are negatives while the y axis is the proportion of predicted negatives that are positive.

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

At x=0.2, the cost of a false positive is more than the cost of a false negative.

(j) [difficult] [MA] 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...

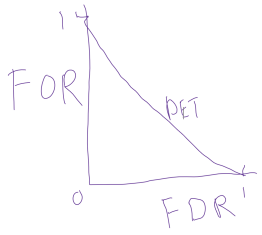

Figure 2: 1.h

The line of random guessing on the ROC curve is the curve x^2 with y-intercept 1 and x-intercept 1.

Problem 2

These are some questions related to 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.

$\text{MSE}(\boldsymbol{x}_*) = \sigma^2 + \text{Bias } (\boldsymbol{x}_*)^2$

(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.

$\text{MSE}(\boldsymbol{x}_*) = \sigma^2 + \text{Bias } [g(\boldsymbol{x}_*)]^2 + \text{Var } [g(\boldsymbol{x}_*)]$

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

$\text{MSE}(\boldsymbol{x}_*) = \sigma^2 + E_x[\text{Bias } [g(\boldsymbol{x}_*)]^2] + E_x[\text{Var } [g(\boldsymbol{x}_*)]]$

(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?

In a, we assumed homoskedascity, so the variance was the same. In b, we assumed random datasets, which led to variance in the δ's.

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

The algorithm is underfit.

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

The algorithm is overfit.

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

There is almost no bias!

(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.

RF reduces MSE by reducing ρ.

(i) [difficult] When can RF lose to bagging M trees? Hint: think hyperparameter choice.

When ρ is too low.

Problem 3

These are some questions related to missingness.

(a) [easy] [MA] What are the three missing data mechanisms? Provide an example when each occurs (i.e., a real world situation). We didn't really cover this in class so I'm making it a MA question only. This concept will NOT be on the exam.

MCAR: There is not specific reason why a certain type of data was not recorded. Say students were given a survey, and certain questions were missing at random (like favorite food, dance, sport)
MAR: There is a specific reason why a certain type of data was not recorded. Say students were given the same survey as above, but only students of Hispanic origin put down a favorite dance.
NMAR: There was a specific survey related reason why a certain type of data was not recorded, i.e. some students amongst themselves decided they will leave food and music blank.

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

It can greatly reduce the amount of data available at hand. In addition, a lot of prediction is possible by using smart methods (imputation).

(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 missing data would be treated as its own category, which would make the prediction more honest.

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

A good default strategy is to use missForest. This is an iterative process and it runs until convergence, using the other data in the training set (which may or may not be missing).

Problem 4

These are some questions related to gradient boosting. The final gradient boosted model after M iterations is denoted G_M which can be written in a number of equivalent ways (see below). The g_t's denote constituent models and the G_t's denote partial sums of the constituent models up to iteration number t. The constituent models are "steps in functional steps" which have a step size η and a direction component denoted $\tilde{g}_t$. The directional component is the base learner $\mathcal{A}$ fit to the negative gradient of the objective function L which measures how close the current predictions are to the real values of the responses:

$$\begin{aligned}
G_M &= G_{M-1} + g_M \\
&= g_0 + g_1 + \ldots + g_M \\
&= g_0 + \eta\tilde{g}_1 + \ldots + \eta\tilde{g}_M \\
&= g_0 + \eta\mathcal{A}\left(\left\langle \boldsymbol{X}, -\nabla L(\boldsymbol{y}, \hat{\boldsymbol{y}}_1)\right\rangle, \mathcal{H}\right) + \ldots + \eta\mathcal{A}\left(\left\langle \boldsymbol{X}, -\nabla L(\boldsymbol{y}, \hat{\boldsymbol{y}}_M)\right\rangle, \mathcal{H}\right) \\
&= g_0 + \eta\mathcal{A}\left(\left\langle \boldsymbol{X}, -\nabla L(\boldsymbol{y}, g_1(\boldsymbol{X}))\right\rangle, \mathcal{H}\right) + \ldots + \eta\mathcal{A}\left(\left\langle \boldsymbol{X}, -\nabla L(\boldsymbol{y}, g_M(\boldsymbol{X}))\right\rangle, \mathcal{H}\right)
\end{aligned}$$

(a) [easy] From a perspective of only multivariable calculus, explain gradient descent and why it's a good idea to find the minimum inputs for an objective function L (in English).

(b) [easy] Write the mathematical steps of gradient boosting for supervised learning below. Use L for the objective function to keep the procedure general. Use notation found in the problem header.

(c) [easy] For regression, what is $g_0(\boldsymbol{x})$?

(d) [easy] For probability estimation for binary response, what is $g_0(\boldsymbol{x})$?

(e) [harder] What are all the hyperparameters of gradient boosting? There are more than just two.

(f) [easy] For regression, rederive the negative gradient of the objective function L.

(g) [easy] For probability estimation for binary response, rederive the negative gradient of the objective function L.

(h) [difficult] For probability estimation for binary response scenarios, what is the unit of the output $G_M(\boldsymbol{x}_\star)$?

(i) [easy] For the base learner algorithm $\mathcal{A}$, why is it a good idea to use shallow CART (which is the recommended default)?

(j) [difficult] For the base learner algorithm $\mathcal{A}$, why is it a bad idea to use deep CART?

(k) [difficult] For the base learner algorithm $\mathcal{A}$, why is it a bad idea to use OLS for regression (or logistic regression for probability estimation for binary response)?

(l) [difficult] If M is very, very large, what is the risk in using gradient boosting even using shallow CART as the base learner (the recommended default)?

(m) [difficult] If η is very, very large but M reasonably correctly chosen, what is the risk in using gradient boosting even using shallow CART as the base learner (the recommended default)?