Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning

ASSIGNMENT 1

Instructor: Purushottam Kar
Date: February 15, 2018
Total: 60 marks

Instructions

1. Only electronic submissions to Gradescope `https://gradescope.com` will be accepted. No hard copy or emailed submissions would be accepted.
2. Your submission should be a single PDF file compiled using the LaTeX style file `saltsubmit.sty`.
3. No zip/tar/png/jpg files or handwritten and then scanned submissions will be accepted.
4. Submissions will be accepted till March 11, 2018, 2359 hrs, IST.
5. Late submissions will be accepted till March 13, 2018, 2359 hrs, IST.
6. Late submissions will incur a penalty – the maximum marks for a late submission will be 80% of the total marks, even if solutions to all questions are absolutely correct.
7. We will be closely checking your submissions for instances of plagiarism.
8. You may be penalized if you do not follow the formatting instructions carefully.
9. Your answer to every question should begin on a new page. The style file is designed to do this automatically. However, if it fails to do so, use the `\clearpage` option in LaTeX before starting the new question, to enforce this.
10. An account has been created for you on Gradescope. Use your IITK CC ID (not GMail, CSE etc IDs) to login and use the "Forgot Password" option to set your password initially.
11. While submitting your assignment on this website, you will have to specify on which page(s) is question 1 answered, on which page(s) is question 2 answered etc. To do this properly, first ensure that the answer to each question starts on a different page.
12. Be careful to flush all your floats (figures, tables) corresponding to question n before starting the answer to question $n+1$ otherwise graders might miss your figures and award you less points. Remember, different people may grade question n and $n+1$.
13. Your solutions must appear in proper order in the PDF file i.e. your solution to question n must be complete in the PDF file (including all plots, tables, proofs etc) before you present a solution to question $n+1$.

Problem 1.1 (Sigmoidal Consistency). Consider a binary classification problem with a distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$. Consider the misclassification loss function $\ell^{0-1}(\hat{y}, y) = \mathbb{I}\{\hat{y} \neq y\}$ for $\hat{y}, y \in \{-1, +1\}$ as well as the sigmoid surrogate loss function $\ell^{\sigma}(\hat{y}, y) = \frac{1}{1+\exp(y\hat{y})}$ for $\hat{y} \in \mathbb{R}$ and $y \in \{-1, +1\}$, which is popularly used to train neural networks. Note that whereas the misclassification loss function expects a proper binary classification as the prediction $\hat{y} \in \{-1, +1\}$, the sigmoid loss can operate with real valued predictions as well i.e. $\hat{y} \in \mathbb{R}$.

For any data point $\mathbf{x} \in \mathcal{X}$, let $\eta(\mathbf{x})$ denote the posterior class probability for the data point $\mathbf{x}$ i.e. $\eta(\mathbf{x}) = \mathbb{P}[Y = 1 \mid X = \mathbf{x}]$. Also, for any $\eta \in [0, 1]$ let $Y \sim \eta$ denote a Rademacher variable that takes value $+1$ with probability η and -1 with probability $1 - \eta$. For any loss function ℓ, and any prediction $\hat{y}$, define

$$L^{\ell}(\hat{y}, \eta) := \mathop{\mathbb{E}}_{Y \sim \eta}[\ell(\hat{y}, Y)],$$

as the expected (conditional) ℓ-loss of predicting $\hat{y}$ on a data point whose true label follows the distribution η. Also denote $L^{\ell}(\eta) := \min_{\hat{y}} L^{\ell}(\hat{y}, \eta)$ as the lowest possible conditional ℓ-loss one can hope to achieve on such a data point using a single best prediction. Define the regret of any predictor f with respect to loss function ℓ as

$$\mathcal{R}^{\ell}_{\mathcal{D}}[f] := \mathop{\mathbb{E}}_{(X,Y) \sim \mathcal{D}}\left[L^{\ell}(f(X), \eta(X)) - L^{\ell}(\eta(X))\right]$$

Show the following results

1. For any $\eta \in [0, 1]$, we have $L^{0-1}(\eta) = L^{\sigma}(\eta) = \min\{\eta, 1 - \eta\}$.
2. For any $\hat{y} \in \{-1, +1\}, \eta \in [0, 1]$, we have $L^{0-1}(\hat{y}, \eta) - L^{0-1}(\eta) = |2\eta - 1| \cdot \mathbb{I}\{\hat{y}(2\eta - 1) < 0\}$.
3. For any $\hat{y} \in \mathbb{R}, \eta \in [0, 1]$, we have $L^{0-1}(\text{sign}(\hat{y}), \eta) - L^{0-1}(\eta) \leq 2\left(L^{\sigma}(\hat{y}, \eta) - L^{\sigma}(\eta)\right)$.
4. Conclude that for any real valued predictor $f \in \mathbb{R}^{\mathcal{X}}$, we have $\mathcal{R}^{0-1}_{\mathcal{D}}[\text{sign} \circ f] \leq 2\mathcal{R}^{\sigma}_{\mathcal{D}}[f]$, where we define $(\text{sign} \circ f)(\mathbf{x}) := \text{sign}(f(\mathbf{x}))$ for any $\mathbf{x} \in \mathcal{X}$.
5. Derive a similar regret transfer bound for the accentuated sigmoid loss $\ell^{\sigma}_{\alpha}(\hat{y}, y) = \frac{1}{1+\exp(\alpha \cdot y\hat{y})}$ for some $\alpha > 0$.

The above shows that achieving small sigmoidal regret automatically ensures small misclassification regret. The sigmoidal loss is easier to work with since it is differentiable but has these nice regret transfer properties as well. Do keep in mind that the minimization over $\hat{y}$ for finding $L^{0-1}(\eta)$ is over Rademacher predictions i.e. over $\hat{y} \in \{-1, +1\}$ whereas the minimization for finding $L^{\sigma}(\eta)$ is over real valued predictions i.e. over $\hat{y} \in \mathbb{R}$. (3+2+3+2+5=15 marks)

Problem 1.2 (Grand Statistical Taxation?). The finance ministry wants to restructure the GST brackets. For this, it wishes to find the κ-th percentile annual expenditure for India for some $\kappa \in (0, 1)$. If we let $\mathcal{I}$ be the set of all Indians where $|\mathcal{I}| =: N$, and for any Indian $i \in \mathcal{I}$, let $E(i)$ denote his/her expenditure, then the finance ministry wants to find a threshold amount v^{κ} such that $\frac{1}{N} \cdot |\{i \in \mathcal{I} : E(i) \leq v^{\kappa}\}| = \kappa$. Since personal expenditure cannot be found from income tax returns (most Indians do not file returns anyway), it seems the only way to do the above is ask each of the $N \approx 1.3 \times 10^9$ Indians what their expenditure is, a daunting task.

However, my expert statistician friend tells me that there is a way out. She instructs me to choose a random set S of $n \ll N$ Indians uniformly at random with replacement, ask them their annual expenditure, and then find an amount $\hat{v}^{\kappa}_S$ such that $\frac{1}{n} \cdot |\{i \in S : E(i) \leq \hat{v}^{\kappa}_S\}| = \kappa$.

Do you agree with the sampling strategy and the construction of the estimator $\hat{v}_S^\kappa$ or would you like to change them? For your choice of strategy and estimator, show that for tolerance and confidence parameters ϵ, δ that the finance ministry desires, you can ensure

$$\underset{S}{\mathbb{P}}\left[\left|\frac{1}{N} \cdot |\{i \in \mathcal{I} : E(i) \leq \hat{v}_S^\kappa\}| - \kappa\right| > \epsilon\right] \leq \delta$$

You may skip small universal constants like $5, 14, 22$ while stating your result but must give explicit dependence on ϵ, δ. You may also assume that $\kappa \cdot N \in \mathbb{N}$ and $\kappa \cdot n \in \mathbb{N}$.

Hint: Do not attempt to show $v^\kappa \approx \hat{v}_S^\kappa$. Such a result may not even be possible (as both v^κ and $\hat{v}_S^\kappa$ are not unique in general) and as such, may or may not guarantee what we require.

Hint: Any threshold v^κ splits the population $\mathcal{I}$ into two groups in whose relative sizes we are interested. (20 marks)

Problem 1.3 (Well ... the search for fairness continues)**.** With the proliferation of learning algorithms in various services, the fear of these algorithms making biased, unfair, or even morally dubious decisions, looms. Consider a labeled domain $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$ over which a regression problem is to be solved with respect to some bounded loss function $\ell : \mathbb{R} \times \mathcal{Y} \to [0, B]$ and an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. Assume that ℓ is L-Lipschitz. I have an algorithm $\mathcal{A} : \bigcup_{n \geq 1} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ that can give, for any training set $S \in (\mathcal{X} \times \mathcal{Y})^n$, a model $\hat{f}_S \in \mathcal{F}$. The algorithm gave me good test performance with respect to ℓ.

However, I am still a bit uneasy since my population $\mathcal{X}$ is actually composed of two equipopulous subpopulations $\mathcal{X}_1$ and $\mathcal{X}_2$. That is to say that $\mathcal{X}_1 \cup \mathcal{X}_2 = \mathcal{X}$, $\mathcal{X}_1 \cap \mathcal{X}_2 = \emptyset$ and $\underset{(X,Y)\sim\mathcal{D}}{\mathbb{P}}[X \in \mathcal{X}_1] = \underset{(X,Y)\sim\mathcal{D}}{\mathbb{P}}[X \in \mathcal{X}_2] = 0.5$. I am worried that $\hat{f}_S$ may be doing a poor job for individuals in subpopulation $\mathcal{X}_2$ but is yet able to shine in terms of overall ℓ-risk by doing extremely well on subpopulation $\mathcal{X}_1$. If this is true then this is a lawsuit waiting to happen (e.g. the subpopulations could be age based and $\hat{f}_S$ could be mispredicting the regression value for old people but getting them accurately for young people and I may have been charging individuals in $\mathcal{X}$ different amounts of money based on the output of the algorithm)

I do not like being sued and want to know if the above is indeed the case. In particular, I am interested in the following notion of fairness for a predictor $f \in \mathcal{F}$

$$\mathsf{FAIR}(f) := \left|\underset{(X,Y)\sim\mathcal{D}}{\mathbb{E}}[\ell(f(X), Y) \,|\, X \in \mathcal{X}_1] - \underset{(X,Y)\sim\mathcal{D}}{\mathbb{E}}[\ell(f(X), Y) \,|\, X \in \mathcal{X}_2]\right|$$

Can you design and analyze an algorithm to estimate $\mathsf{FAIR}(\hat{f}_S)$? You may assume that the model class $\mathcal{F}$ is nice in terms of having "small" Rademacher complexity or VC dimension or covering number, depending on your analysis technique. However, do note that the algorithm $\mathcal{A}$ was not designed to optimize any notion of fairness, including FAIR. Is it possible to estimate $\mathsf{FAIR}(\hat{f}_S)$ using S itself? (25 marks)