CIS 520: Machine Learning *Due:* Mar. 23, 2021 (11:59 AM EDT)

Problem Set 3

Submission Instructions:

This is your main submission, and must contain your solutions to **all problems that you are attempting**. It must be submitted as a **s**ingle PDF file to **Gradescope**, compiled in LATEX using the LATEX template provided on Canvas. Each problem should be on a new page. Mathematical problems can be either typed in LATEX or written by hand and **scanned** into images included in your LATEX solution, but it must be **readable** by our staff to be graded; we recommend that you not take photos of your hand-written solutions.

Special instruction for submitting on Gradescope: For each problem, select the pages containing your solution for that problem.

Summary: The PDF file should be submitted to the corresponding assignment on Gradescope. All components of problem set must be received in the right places and in the right formats before the submission deadline. Plan to submit early!

Collaboration Policy:

You are allowed to discuss problems in groups, but you must write all your solutions and code **individually, on your own**. All collaborators with whom problems were discussed **must** be listed in your PDF submission.

1. (20 points) **Decision Trees.** Consider a binary classification task on $\mathcal{X} = \mathbb{R}^2$ with the following set of 8 training examples:

i	$\mathbf{x}_i$	y_i
1	$(3, 9)$	$+1$
2	$(7, 11)$	-1
3	$(4, 6)$	$+1$
4	$(7, 2)$	$+1$
5	$(9, 9)$	-1
6	$(10, 2)$	$+1$
7	$(6, 12)$	-1
8	$(6, 1)$	$+1$

You would like to construct a decision tree classifier based on the above training examples and are trying to decide what split to use at the root node. Suppose you are considering the following two splits: (1) $x_1 > 5$ and (2) $x_2 > 8$.

(a) What is the entropy associated with a single leaf node containing all the above examples? What is the Gini index associated with such a leaf node? Show your calculations.

(b) For each of the above two splits, calculate the information gain that would result from using that split at the root node. Show your calculations. Which of these splits would you choose based on the entropy criterion?

(c) For each of the above two splits, calculate the Gini reduction that would result from using that split at the root node. Show your calculations. Which of these splits would you choose based on the Gini index criterion?

2. (7 points) **Choice of α_t in AdaBoost.** Recall that on each round t, the AdaBoost algorithm receives a weak classifier h_t with D_t-weighted error er_t, sets

$$\alpha_t \leftarrow \frac{1}{2} \ln \left(\frac{1 - \text{er}_t}{\text{er}_t} \right),$$

and then updates the weights for the next round according to

$$D_{t+1}(i) \leftarrow \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

with normalization factor Z_t given by

$$Z_t = \sum_{j=1}^{m} D_t(j) \exp(-\alpha_t y_j h_t(x_j)) \,.$$

Show that the above choice of α_t minimizes Z_t.

(Hint: Treat Z_t as a function of α_t and set the derivative with respect to α_t to zero to obtain a stationary point; then check the sign of the second derivative at this stationary point.)

3. (18 points) **Multiclass Boosting.** In this problem you will analyze the AdaBoost.M1 algorithm, a multiclass extension of AdaBoost. Given a training sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$, where x_i are instances in some instance space $\mathcal{X}$ and y_i are multiclass labels that take values in $\{1, \ldots, K\}$, the algorithm maintains weights $D_t(i)$ over the examples (x_i, y_i) as in AdaBoost, and on round t, gives the weighted sample (S, D_t) to the weak learner. The weak learner returns a multiclass classifier $h_t : \mathcal{X} \rightarrow \{1, \ldots, K\}$ with weighted error less than $\frac{1}{2}$; here the weighted error of h_t is measured as

$$\text{er}_t = \sum_{i=1}^{m} D_t(i) \cdot \mathbf{1}(h_t(x_i) \neq y_i) \,.$$

Note that the assumption on the weak classifiers is stronger here than in the binary case, since we require the weak classifiers to do more than simply improve upon random guessing (there are other multiclass boosting algorithms that allow for weaker classifiers; you will analyze the simplest case here). For convenience, we will encode the weak classifier h_t as $\widetilde{h}_t : \mathcal{X} \to \{\pm 1\}^K$, where

$$\widetilde{h}_{t,k}(x) = \begin{cases} +1 & \text{if } h_t(x) = k \\ -1 & \text{otherwise.} \end{cases}$$

In other words, $\widetilde{h}_t(x)$ is a K-dimensional vector that contains $+1$ in the position of the predicted class for x and -1 in all other $(K-1)$ positions. On each round, AdaBoost.M1 re-weights examples such that examples misclassified by the current weak classifier receive higher weight in the next round. At the end, the algorithm combines the weak classifiers h_t via a weighted majority vote to produce a final multiclass classifier H:

```
Algorithm AdaBoost.M1
Inputs: Training sample $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \{1, \ldots, K\})^m$
        Number of iterations $T$
Initialize: $D_1(i) = \frac{1}{m} \quad \forall i \in [m]$
For $t = 1, \ldots, T$:
  – Train weak learner on weighted sample $(S, D_t)$; get weak classifier $h_t : \mathcal{X} \to \{1, \ldots, K\}$
  – Set $\alpha_t \leftarrow \frac{1}{2} \ln\left(\frac{1 - \mathrm{er}_t}{\mathrm{er}_t}\right)$
  – Update:
      $D_{t+1}(i) \leftarrow \frac{D_t(i) \exp(-\alpha_t \widetilde{h}_{t,y_i}(x_i))}{Z_t}$
    where $Z_t = \sum_{j=1}^m D_t(j) \exp(-\alpha_t \widetilde{h}_{t,y_j}(x_j))$
Output final hypothesis:
    $H(x) \in \arg\max_{k \in \{1,\ldots,K\}} \underbrace{\sum_{t=1}^T \alpha_t \widetilde{h}_{t,k}(x)}_{F_{T,k}(x)}$
```

You will show, in five parts below, that if all the weak classifiers have error er_t at most $\frac{1}{2} - \gamma$, then after T rounds, the training error of the final classifier H, given by

$$\mathrm{er}_S[H] = \frac{1}{m} \sum_{i=1}^m \mathbf{1}(H(x_i) \neq y_i) \,,$$

is at most $e^{-2T\gamma^2}$ (which means that for large enough T, the final error $\mathrm{er}_S[H]$ can be made as small as desired).

(a) Show that

$$D_{T+1}(i) \; = \; \frac{\frac{1}{m} \, e^{-F_{T,y_i}(x_i)}}{\prod_{t=1}^T Z_t} \,.$$

(b) Show that

$$\mathbf{1}(H(x_i) \neq y_i) \;\leq\; \mathbf{1}\big(F_{T,y_i}(x_i) \leq 0\big) \,.$$

(*Hint:* Consider separately the two cases $H(x_i) \neq y_i$ and $H(x_i) = y_i$, and note that $\sum_{k=1}^K F_{T,k}(x_i) = -(K-2) \sum_{t=1}^T \alpha_t$.)

(c) Show that

$$\mathrm{er}_S[H] \;\le\; \frac{1}{m}\sum_{i=1}^{m} e^{-F_{T,y_i}(x_i)} \;=\; \prod_{t=1}^{T} Z_t\,.$$

(*Hint:* For the inequality, use the result of part (b) above, and the fact that $\mathbf{1}(u \le 0) \le e^{-u}$; for the equality, use the result of part (a) above.)

(d) Show that for the given choice of α_t, we have

$$Z_t \;=\; 2\sqrt{\mathrm{er}_t(1-\mathrm{er}_t)}\,.$$

(e) Suppose $\mathrm{er}_t \le \frac{1}{2} - \gamma$ for all t (where $0 < \gamma \le \frac{1}{2}$). Then show that

$$\mathrm{er}_S[H] \;\le\; e^{-2T\gamma^2}\,.$$

4. (10 points) **Multiclass Classification with a Cost-Sensitive Loss.** You are collaborating with a cancer treatment center and are trying to help them predict which patients will respond well to a particular cancer drug. You are given clinical data for patients they have given the drug to in the past; for each such patient, the data contains measurements from the patient's tumor biopsy together with a class label indicating whether the patient was a complete responder (CR) to the drug, a partial responder (PR), or a non-responder (NR). Your goal is to predict the response category (CR, PR, or NR) for new patients based on their tumor biopsy measurements. You are given the following loss for this problem:

		$\widehat{y}$		
		NR	PR	CR
y	NR	0	5	4
	PR	9	0	1
	CR	10	1	0

Here mis-predicting a PR case as CR or vice versa incurs relatively little cost, since in both cases the patient is given the drug. Mis-predicting a PR or CR case as NR is very costly, since in this case a patient who could benefit from the drug does not receive treatment. Mis-predicting an NR case as PR or CR is also costly, since it involves unnecessary expense and side effects for a patient who doesn't benefit from the drug, but is less costly than errors in the other direction.

Suppose that, using the training data provided to you, you have trained a CPE model, and that for 2 new patients with measurement vectors x_1 and x_2, the model estimates the following probabilities for the 3 classes:

$$\text{Patient 1:} \quad \begin{pmatrix} \widehat{\eta}_{\mathrm{NR}}(x_1) \\ \widehat{\eta}_{\mathrm{PR}}(x_1) \\ \widehat{\eta}_{\mathrm{CR}}(x_1) \end{pmatrix} = \begin{pmatrix} 0.6 \\ 0.35 \\ 0.05 \end{pmatrix}; \qquad \text{Patient 2:} \quad \begin{pmatrix} \widehat{\eta}_{\mathrm{NR}}(x_2) \\ \widehat{\eta}_{\mathrm{PR}}(x_2) \\ \widehat{\eta}_{\mathrm{CR}}(x_2) \end{pmatrix} = \begin{pmatrix} 0.2 \\ 0.45 \\ 0.35 \end{pmatrix}.$$

Which response category would you predict for each patient, and how do these categories differ from what you would have predicted under the 0-1 loss? Explain your answers.

5. (15 points) **Performance Measures for Face Detection in Images.** Consider a face detection problem, where the goal is to build a system that can automatically detect faces in images. Two research groups develop systems for this problem using slightly different approaches. In both cases, the central component of the system is a binary classifier which, when applied to a 24 × 24 image, decides whether or not it is a face. The two groups train their classifiers using different learning algorithms. Moreover, when given a new image, they also apply their classifiers in slightly different ways: group A tests 24 × 24 regions of the image taking strides of size 2 (so, for example, for a 100 × 100 image, $(39)^2$

regions would be tested); group B tests 24 × 24 regions of the image taking strides of size 5 (so here, for a 100 × 100 image, only $(16)^2$ regions would be tested).[1] On a standard benchmark suite of test images that contains 300 faces altogether, the two groups have the following performances (assume the regions tested by both systems include all the 300 true face regions):

Research group	Number of regions tested	Number of faces detected correctly	Number of non-face regions detected as faces
A	24,000	260	100
B	15,000	245	65

(a) Based on the above numbers, calculate the TPR (recall), TNR, and precision of each group's system as tested. Also calculate the resulting geometric mean (GM) and F_1 measures for each system. If you were to select a system based on the GM measure, which system would you choose? Would your choice change if you were to select a system based on the F_1 measure?

(b) Which performance measure would be more suitable for this problem – the GM measure or the F_1 measure? Why?

6. (12 points) **VC-Dimension Based Generalization Error Bounds.** Recall that for any binary classification algorithm which given a training sample $S \in (\mathcal{X} \times \{\pm 1\})^m$ returns a binary classifier $h_S : \mathcal{X} \to \{\pm 1\}$ from a function class $\mathcal{H}$ of finite VC-dimension, and for any distribution D on $\mathcal{X} \times \{\pm 1\}$ and confidence parameter $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the draw of $S \sim D^m$, the generalization error of the learned function h_S can be upper bounded in terms of its training error as follows:

$$\mathrm{er}_D^{0\text{-}1}[h_S] \;\leq\; \widehat{\mathrm{er}}_S^{0\text{-}1}[h_S] + \sqrt{\frac{8\Big(\mathrm{VCdim}(\mathcal{H}) \cdot (\ln(2m) + 1) + \ln(\frac{4}{\delta})\Big)}{m}}\,.$$

Let $\mathcal{X} = \mathbb{R}^2$, and suppose that given a training sample S, you use the SVM algorithm to learn a linear classifier h_S^1 and kernel-based classifiers h_S^2, h_S^3 with polynomial kernels of degrees 2 and 3, respectively.

(a) Suppose you are given a training sample S containing 2000 examples; assume all examples are drawn iid from some unknown distribution D. Say you learn three classifiers as above and observe they have the following training errors:

$$\widehat{\mathrm{er}}_S^{0\text{-}1}[h_S^1] = 0.18\,; \quad \widehat{\mathrm{er}}_S^{0\text{-}1}[h_S^2] = 0.13\,; \quad \widehat{\mathrm{er}}_S^{0\text{-}1}[h_S^3] = 0.08\,.$$

Compute high confidence upper bounds on the generalization error of the three classifiers, using confidence parameter $\delta = 0.01$. Show your calculations. If you were to select one of these classifiers based on the training errors, which classifier would you select? If you were to select a classifier using the 99% confidence generalization error bounds you have derived, which classifier would you select?[2]

(b) Now suppose you are given a training sample S containing 20,000 examples; again, assume all examples are drawn iid from some unknown distribution D. Say you learn three classifiers as above and observe they have the same training errors as in part (a):

$$\widehat{\mathrm{er}}_S^{0\text{-}1}[h_S^1] = 0.18\,; \quad \widehat{\mathrm{er}}_S^{0\text{-}1}[h_S^2] = 0.13\,; \quad \widehat{\mathrm{er}}_S^{0\text{-}1}[h_S^3] = 0.08\,.$$

Repeat the calculations of part (a). If you were to select one of these classifiers based on the training errors, which classifier would you select? If you were to select a classifier using the 99% confidence generalization error bounds you have derived, which classifier would you select?

[1] In practice, the 24 × 24 classifier would also be applied to multiple scaled versions of the input image; we ignore this issue here for simplicity.

[2] Note that, technically, to compare the three bounds, we need all three bounds to hold simultaneously; since the probability of any of them failing individually is at most 0.01, using the union bound, the probability of any one of them failing is at most 0.03, and therefore the bounds hold simultaneously with probability at least 0.97. (Alternatively, if we want them to hold simultaneously with probability at least 0.99, we can compute the individual bounds at confidence level $\delta = 0.01/3$ each).

7. (18 points) **Bias-Variance Decomposition for Binary Class Probability Estimation.** In this problem, you will prove a decomposition for the expected cross-entropy loss for binary class probability estimation (CPE), analogous to the bias-variance decomposition for the expected squared loss for regression. Specifically, consider a binary CPE problem with some instance space $\mathcal{X}$ and label space $\mathcal{Y} = \{\pm 1\}$, and let D be an unknown probability distribution on $\mathcal{X} \times \mathcal{Y}$ from which labeled examples are generated. Say you have an algorithm which, given a training sample S, produces a CPE model $\widehat{\eta}_S : \mathcal{X} \to [0,1]$. Recall that the cross-entropy loss of such a model $\widehat{\eta}_S$ on a new example (x, y) is given by

$$\ell_{\text{CE}}(y, \widehat{\eta}_S(x)) = -\mathbf{1}(y = +1) \cdot \ln(\widehat{\eta}_S(x)) - \mathbf{1}(y = -1) \cdot \ln(1 - \widehat{\eta}_S(x))\,,$$

and the corresponding generalization error is then

$$\begin{aligned} \text{er}_D^{\text{CE}}[\widehat{\eta}_S] &= \mathbf{E}_{(X,Y)\sim D}[\ell_{\text{CE}}(Y, \widehat{\eta}_S(X))] \\ &= \mathbf{E}_X\left[\mathbf{E}_{Y|X}[\ell_{\text{CE}}(Y, \widehat{\eta}_S(X))]\right] \\ &= \mathbf{E}_X\big[-\eta(X)\,\ln(\widehat{\eta}_S(X)) - (1-\eta(X))\,\ln(1-\widehat{\eta}_S(X))\big]\,. \end{aligned}$$

We are interested in understanding the *expected* generalization error over multiple random training samples S:

$$\mathbf{E}_S\big[\text{er}_D^{\text{CE}}[\widehat{\eta}_S]\big] \;=\; \mathbf{E}_X\big[\mathbf{E}_S\big[-\eta(X)\,\ln(\widehat{\eta}_S(X)) - (1-\eta(X))\,\ln(1-\widehat{\eta}_S(X))\big]\big]\,.$$

We start with some information-theoretic preliminaries that you will need for this problem.

- **Entropy.** The entropy of a probability distribution measures the amount of 'randomness' in the distribution. The entropy of a Bernoulli distribution with parameter $p \in [0,1]$ is given by

$$H(p) = -p \ln p - (1-p)\ln(1-p)\,.$$

- **Cross-entropy.** The cross-entropy from one probability distribution to another measures roughly how bad it is to use the first distribution in place of the second. The cross-entropy from a Bernoulli distribution with parameter q (often an 'estimated' distribution) to another Bernoulli distribution with parameter p (often a 'true' distribution) is given by

$$H(p,q) = -p \ln q - (1-p)\ln(1-q)\,.$$

- **Kullback-Leibler (KL) divergence.** The KL divergence (also known as relative entropy) is a form of (asymmetric) distance between probability distributions. As with the cross-entropy, the KL divergence from one probability distribution to another also measures roughly how bad it is to use the first distribution in place of the second (indeed, it is related to the cross-entropy simply via subtraction of an entropy term). The KL divergence from a Bernoulli distribution with parameter q (often an 'estimated' distribution) to another Bernoulli distribution with parameter p (often a 'true' distribution) is given by

$$\begin{aligned} \text{KL}(p\,\|\,q) &= p\ln\left(\frac{p}{q}\right) + (1-p)\ln\left(\frac{1-p}{1-q}\right) \\ &= H(p,q) - H(p)\,. \end{aligned}$$

  This has the property that $\text{KL}(p\,\|\,p) = 0$.

- **'Average' model under KL divergence.** In the case of regression under squared loss, an 'average' regression model is given by $\bar{f}(x) = \mathbf{E}_S[f_S(x)]$. This has the property that it minimizes the expected squared loss: $\bar{f}(x) \in \arg\min_{c\in\mathbb{R}} \mathbf{E}_S[(f_S(x) - c)^2]$. In the case of binary CPE under cross-entropy loss, we will consider an 'average' CPE model given by

$$\bar{\eta}(x) = \frac{1}{Z(x)} e^{\mathbf{E}_S[\ln \widehat{\eta}_S(x)]}\,,$$

  where $Z(x) = e^{\mathbf{E}_S[\ln \widehat{\eta}_S(x)]} + e^{\mathbf{E}_S[\ln(1-\widehat{\eta}_S(x))]}$. This has a similar property as in the regression case, namely, that $\bar{\eta}(x) \in \arg\min_{c\in[0,1]} \mathbf{E}_S\big[\text{KL}(c\,\|\,\widehat{\eta}_S(x))\big]$.

(a) Using the above preliminaries, show that the expected generalization error can be decomposed as

$$\mathbf{E}_S\big[\mathrm{er}_D^{\mathrm{CE}}[\widehat{\eta}_S]\big] \quad = \quad \underbrace{\mathbf{E}_X\big[\mathbf{E}_S\big[\mathrm{KL}(\bar{\eta}(X)\,\|\,\widehat{\eta}_S(X))\big]\big]}_{\text{term 1}} + \underbrace{\mathbf{E}_X\big[\mathrm{KL}(\eta(X)\,\|\,\bar{\eta}(X))\big]}_{\text{term 2}} + \underbrace{\mathbf{E}_X\big[H(\eta(X))\big]}_{\text{term 3}}\,.$$

(b) Give an interpretation for each of the three terms in the above decomposition and explain how they play an analogous role to the terms in the standard bias-variance decomposition of the expected squared loss for regression.

(Hints for part (a): Show the decomposition for a fixed instance x first, and then take expectations over x. Start by showing that $\mathrm{KL}(\eta(x)\,\|\,\bar{\eta}(x)) = \mathbf{E}_S[\mathrm{KL}(\eta(x)\,\|\,\widehat{\eta}_S(x))] + \ln Z(x)$. *Then show that* $\ln Z(x) = -\mathbf{E}_S[\mathrm{KL}(\bar{\eta}(x)\,\|\,\widehat{\eta}_S(x))]$*; putting everything together should then give the result.)*