CS 189 Introduction to Machine Learning
Spring 2023 Jonathan Shewchuk

DIS13

1 VC Dimension of Half-Spaces

In lecture, we discussed how the VC dimension of linear classifiers in a plane is 3. In this problem, we will shorten our set of hypotheses to only be those linear classifiers that pass through the origin, and provide a rigorous proof of how the VC dimension of such classifiers in $\mathbb{R}^d$ is d.

We will define a linear classifier as a function $f_w \to \{-1, 1\}$ such that

$$f_w(x) = \text{sign}\left(w^T x\right)$$

and the class of half-spaces of dimension d to be defined as

$$\mathcal{H}_d = \{f_w \ : \ w \in \mathbb{R}^d\}$$

(a) Show that there exists a set of points $S \in \mathbb{R}^d$ of size d such that $\mathcal{H}_d$ shatters S.

(b) We now define a set $T = \{x^{(1)} \dots x^{(d+1)}\}$ of size $d + 1$. Find two sets $I, J \subseteq T$ where $I \cap J = \emptyset$ and with *at least* one of I or J nonempty, as well as positive coefficients $a_1 \dots a_{d+1}$ such that

$$\sum_{i \in I} a_i x^{(i)} = \sum_{j \in J} a_j x^{(j)}$$

(c) Using the last two parts, prove that $\text{VC}(\mathcal{H}_d) = d$

Hint: use a proof by contradiction by assuming that $\mathcal{H}_d$ shatters T and use the linearity of the inner product

2 Boosted Decision Trees

In this problem, we'll develop the key concepts required to understand boosted decision trees using the AdaBoost algorithm. We are given data $D = \{(X_i, y_i)\}_{i=1}^N$, where $X_i \in \mathbb{R}^d$ and $y_i \in C = \{-1, 1\}$. Recall that, for a node in a decision tree with data $S \subseteq D$, we compute the proportion of each label, then use those proportions to compute the entropy:

$$p_c = \frac{1}{|S|} \sum_{(X,y)\in S} I(y_i = c),$$

$$H(S) = \sum_{c\in C} -p_c \ln p_c.$$

(a) Let w_i be the **weight** of observation (X_i, y_i). The weight of an observation can be thought of as its importance. To incorporate weights into our decision tree, we redefine the way we compute proportions. Let $Z = \sum_{i=1}^{|S|} w_i$, and

$$p_c = \frac{1}{Z} \sum_{i=1}^{|S|} I(y_i = c)\, w_i.$$

Assume $w_i = a$ for all i. Show that $H(S)$ does not change for constant values of a.

(b) Like Random Forest, boosting is an ensemble method. We train several decision trees on weighted observations and combine their predictions to construct an overall improved classifier. The Adaboost algorithm starts by training the first decision tree G_1 using observation weights initialized to $w_i = \frac{1}{|S|}$. The weighted error rate of a trained tree G_t is given by

$$\text{err}_t = \frac{\sum_{i=1}^{|S|} w_i I(y_i \neq G_t(X_i))}{\sum_{i=1}^{|S|} w_i}.$$

Each tree G_t in a boosted ensemble is assigned a weight. The weight is computed using the negative logit function (you will show this is optimal in the lecture on Boosting)

$$\beta_t = \frac{1}{2} \ln\left(\frac{1 - \text{err}_t}{\text{err}_t}\right).$$

What are the minimum and maximum possible values, $\text{err}_{\min}$ and $\text{err}_{\max}$, of err_t so that $\text{err}_{\min} \leq \text{err}_t \leq \text{err}_{\max}$?

(c) After training T decision trees, the AdaBoost algorithm produces the following decision function

$$G(x) = \text{sign}\left[\sum_{t=1}^{T} \beta_t G_t(x)\right],$$

where $\text{sign}[x] = 1$ if $x \geq 0$, and -1 otherwise. Compute $\nabla_{\text{err}_t} \beta_t$. What do you notice about the rate of change of β_t when err_t is near its bounds?

(d) After each decision tree is trained, the weight for each observation $i = 1, ..., |S|$ is updated by the following update rule:

$$w_i \leftarrow w_i \exp\left(-\beta_t y_i G_t(x_i)\right).$$

Note that $-y_i G_t(x_i) = 2I\left(y_i \neq G_t x_i\right) - 1$. Since the -1 becomes a multiplicative constant, we can drop it to obtain

$$w_i \leftarrow w_i \exp\left(2\beta_t I\left(y_i \neq G_t x_i\right)\right).$$

The subsequent tree is trained on these updated weights.

(a) What is the value of $w_j^{(2)}$ if observation j is the only observation that has not been classified correctly after 1 iteration?

(b) How does this influence the optimal split choice for nodes in decision tree G_2?

(c) What is $w_j^{(3)}$ if j is still the only observation which has not been classified correctly?

3 Cal vs. Stanford Decision Stumps

Recall that **ensembling** is the practice of training several models to perform the same task. A random forest is an example of an ensemble, particularly of decision trees. Also recall that **boosting** is the practice of iteratively training decision trees that learn from the errors of the previous trees. **AdaBoost** is one such boosting algorithm.

In this example, we deal with decision stumps, which are one-level decision trees. AdaBoost is often used to ensemble decision stumps, which we will explore in this problem.

Recall that in AdaBoost, our input is an $n \times d$ design matrix X with n labels $y_i = \pm 1$, and at the end of iteration T the importance of each sample is reweighted as

$$w_i^{(T+1)} = w_i^{(T)} \exp(-\beta_T \, y_i \, G_T(X_i)), \quad \text{where} \quad \beta_T = \frac{1}{2} \ln\left(\frac{1 - \text{err}_T}{\text{err}_T}\right) \quad \text{and} \quad \text{err}_T = \frac{\sum_{y_i \neq G_T(X_i)} w_i^{(T)}}{\sum_{i=1}^{n} w_i^{(T)}}.$$

Note that err_T is the weighted error rate of the classifier G_T. Recall that $G_T(z)$ is ± 1 for all points z, but the metalearner has a non-binary decision function $M(z) = \sum_{t=1}^{T} \beta_t G_t(z)$. To classify a test point z, we calculate $M(z)$ and return its sign.

We went to use AdaBoost to train an ensemble of decision stumps to classify whether a student goes to Stanford and Cal, based on their fitness activity. Our training data is shown below, where +1 corresponds to Cal and -1 corresponds to Stanford:

Bike Miles Driven	Elevation Climbed	Cal or Stanford
2	1	+1
4	2	-1
5	5	+1

(a) Write out the decision function for $G_t(X_i)$, which is decision stump t classifying a point X_i. Assume that for this decision stump, the entropy-minimizing feature is feature j, so we use only the scalar value X_{ij} in classification. Assume also that for this entropy-minimizing feature, we split on the threshold α_t.

(b) Assume that decision stump G_i has $\texttt{err}_i \geq 0.5$. What is the range of the resultant value of β_i? What is the significance of this range for a classified test point?

(c) Given the dataset, find decision stump G_1, model error $\texttt{err}_1$, and model weight β_1. Assume that the weights are initialized to $w_1^{(0)} = w_2^{(0)} = w_3^{(0)} = \frac{1}{3}$. (Hint: many thresholds will achieve the lowest loss, so let's use Bike Miles Driven > 3 for the decision stump.)

(d) Compute the weights of the points after the creation of decision stump G_1. In other words, compute $w_1^{(1)}$, $w_2^{(1)}$, and $w_3^{(1)}$.

(e) Find decision stump G_2, model error $\texttt{err}_2$, and model weight β_2. (Hint: many thresholds will achieve the lowest loss, so let's use Elevation Climbed > 3.5 for the decision stump.)

(f) Compute the weights of the points after the creation of decision stump G_2. In other words, compute $w_1^{(2)}$, $w_2^{(2)}$, and $w_3^{(2)}$. Which point do you think decision stump G_3 will prioritize for correct classification?

(g) We halt training after two iterations, i.e. our forest has 2 decision stumps. Classify the following test point: (7, 4).

