M2608.001300
Machine Learning Fundamentals & Applications

Assignment #2
(100 points total)

Due 11:59pm on April 22, 2019

SHOULD BE DONE AND SUBMITTED INDIVIDUALLY (NO TEAM ASSIGNMENT)

Problem 1 (20 Points)

This problem investigates how changing the error measure can change the result of the learning process. You have N data points $y_1 \leq \cdots \leq y_N$ and wish to estimate a 'representative' value.

(a) [10 pts] If your algorithm is to find the hypothesis h that minimizes the in-sample sum of *squared deviations*,

$$\mathrm{E}_{\mathrm{in}}(h) = \sum_{n=1}^{N} (h - y_n)^2,$$

then show that your estimate will be the in sample mean,

$$h_{\mathrm{mean}} = \frac{1}{N} \sum_{n=1}^{N} y_n.$$

(b) [10 pts] If your algorithm is to find the hypothesis h that minimizes the in-sample sum of *absolute deviations*,

$$\mathrm{E}_{\mathrm{in}}(h) = \sum_{n=1}^{N} |h - y_n|,$$

then show that your estimate will be the in-sample median h_{med}, which is any value for which half the data points are at most h_{med} and half the data points are at least h_{med}.

Problem 2 (20 Points)

Consider

$$\mathsf{e}_n(\mathbf{w}) = \max(0, 1 - y_n \mathbf{w}^\top \mathbf{x}_n).$$

(a) [5 pts] Show that $\mathsf{e}_n(\mathbf{w})$ is continuous and differentiable except when $y_n = \mathbf{w}^\top \mathbf{x}_n$.

(b) [5 pts] Show that $\mathsf{e}_n(\mathbf{w})$ is an upper bound for the "0-1 loss" or $[\![\text{sign}(\mathbf{w}^\top \mathbf{x}_n) \neq y_n]\!]$. Thus, $\frac{1}{N}\sum_{n=1}^{N} \mathsf{e}_n(\mathbf{w})$ is an upper bound for the in-sample classification error $\mathrm{E}_{\text{in}}(\mathbf{w})$.

(c) [10 pts] Apply stochastic gradient descent on $\frac{1}{N}\sum_{n=1}^{N} \mathsf{e}_n(\mathbf{w})$ (ignoring the singular case of $\mathbf{w}^\top \mathbf{x}_n = y_n$) and derive a new perceptron learning algorithm.

Note: $\mathsf{e}_n(\mathbf{w})$ corresponds to the "hinge loss" used for maximum-margin classification, most notably for support vector machines (SVMs).

Problem 3 (20 Points)

There are a number of bounds on the generalization error ϵ, all holding with probability at least $1-\delta$.

(a) Original VC-bound:

$$\epsilon \leq \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}$$

(b) Rademacher penalty bound:

$$\epsilon \leq \sqrt{\frac{2\ln(2Nm_{\mathcal{H}}(N))}{N}} + \sqrt{\frac{2}{N} \ln \frac{1}{\delta}} + \frac{1}{N}$$

(c) Parrondo and van den Broek:

$$\epsilon \leq \sqrt{\frac{1}{N}\left(2\epsilon + \ln \frac{6m_{\mathcal{H}}(2N)}{\delta}\right)}$$

(d) Devroye:

$$\epsilon \leq \sqrt{\frac{1}{2N}\left(4\epsilon(1+\epsilon) + \ln \frac{4m_{\mathcal{H}}(N^2)}{\delta}\right)}$$

Note that (c) and (d) are implicit bounds in ϵ. Fix $d_{\text{VC}} = 50$ and $\delta = 0.05$. Plot these bounds as a function of N. Which is the best?

Problem 4 (20 Points)

The bias-variance decomposition of out-of-sample error is based on squared error measures. Recall that the out-of-sample error is given by

$$\mathrm{E}_{\text{out}}\left(g^{(\mathcal{D})}\right) = \mathbb{E}_{\mathbf{x}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right] \tag{1}$$

where $\mathbb{E}_{\mathbf{x}}$ denotes the expected value with respect to input $\mathbf{x}$, and using $g^{(\mathcal{D})}$ is to make explicit the dependence of g on data $\mathcal{D}$. From (1) we can remove the dependence on $\mathcal{D}$ by taking average:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[\mathrm{E}_{\text{out}}\left(g^{(\mathcal{D})}\right)\right] &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{x}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right]\right] \\ &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right]\right].\end{aligned}$$

(a) [5 pts] To evaluate $\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right]$, we define the 'average' hypothesis $\overline{g}(\mathbf{x})$ as

$$\overline{g}(\mathbf{x}) \triangleq \mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\mathbf{x})\right]. \tag{2}$$

Now imagine that we have K datasets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$, then what will be the average hypothesis $\overline{g}(\mathbf{x})$ estimated using these datasets?

(b) [10 pts] Let us define

$$\begin{aligned}\mathsf{var}(\mathbf{x}) &\triangleq \mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \overline{g}(\mathbf{x})\right)^2\right], \text{ and} \\ \mathsf{bias}(\mathbf{x}) &\triangleq \left(\overline{g}(\mathbf{x}) - f(\mathbf{x})\right)^2.\end{aligned}$$

Show that

$$\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right] = \mathsf{var}(\mathbf{x}) + \mathsf{bias}(\mathbf{x})$$

(c) [5 pts] Show that

$$\mathbb{E}_{\mathcal{D}}\left[\mathrm{E}_{\text{out}}(g^{(\mathcal{D})})\right] = \mathsf{bias} + \mathsf{var}$$

where $\mathsf{bias} \triangleq \mathbb{E}_{\mathbf{x}}\left[\mathsf{bias}(\mathbf{x})\right]$ and $\mathsf{var} \triangleq \mathbb{E}_{\mathbf{x}}\left[\mathsf{var}(\mathbf{x})\right]$.

Problem 5 (20 Points)

In support vector machines, the hyperplane h separates the data if and only if it can be represented by weights $(b, \mathbf{w})$ that satisfy

$$\min_{n=1,\ldots,N} y_n(\mathbf{w}^\top \mathbf{x}_n + b) = 1. \tag{3}$$

Consider the data below and a 'hyperplane' $(b, \mathbf{w})$ that separates the data.

$$\mathbf{X} = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ +1 \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} 1.2 \\ -3.2 \end{bmatrix} \qquad b = -0.5$$

(a) [10 pts] Compute

$$\rho = \min_{1,\ldots,N} y_n(\mathbf{w}^\top \mathbf{x}_n + b).$$

(b) [5 pts] Compute the weights $\frac{1}{\rho}(b, \mathbf{w})$ and show that they satisfy Eq (3).

(c) [5 pts] Plot both hyperplanes to show that they are the *same* separator.

Note: This problem is to give a concrete example of re-normalizing the weights to show that the condition (3) and the following condition are equivalent, as covered at class:

$$y_n(\mathbf{w}^\top \mathbf{x}_n + b) > 0. \tag{4}$$