ELEG/CISC 867 Advanced Machine Learning

Homework 1 University of Delaware
Handout: March 12, 2019 Due: March 21, 2019

PROBLEM 1 (BAYES PREDICTOR; 30 PTS). Suppose $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ are jointly distributed according to distribution P. For any predictor $h : \mathcal{X} \to \hat{\mathcal{Y}}$ that maps input X to predicted output $\hat{Y}$, define the risk of h under distribution P and loss function ℓ as

$$L(h, P) \triangleq \mathbb{E}_P[\ell(Y, h(X))].$$

The predictor

$$f = \operatorname*{argmin}_h L(h, P)$$

that minimizes the risk is called the *Bayes predictor*, and its resultant risk

$$\min_h L(h, P)$$

is called the *Bayes risk*. Show that under different loss functions, the Bayes predictor takes different forms as follows and derive their resultant Bayes risks.

(a) 0-1 loss: Show that under the 0-1 loss, the Bayes predictor f is given by the well-known maximum a posteriori (MAP) rule, i.e.,

$$f(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} p_{Y|X}(y|x), \quad \text{[5 pts]}$$

with the Bayes risk

$$L(f, P) = 1 - \sum_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} p_{X,Y}(x, y). \quad \text{[5 pts]}$$

(b) Square loss: Show that under the square loss, the Bayes predictor f is given by the conditional expectation of Y given $X = x$, i.e.,

$$f(x) = \mathbb{E}_P[Y|X = x], \quad \text{[5 pts]}$$

with the Bayes risk

$$L(f, P) = \mathbb{E}_P[\mathrm{Var}(Y|X)]. \quad \text{[5 pts]}$$

(c) Log loss: Show that under the log loss, the Bayes predictor f is given by the conditional distribution of Y given $X = x$, i.e.,

$$[f(x)](y) = p_{Y|X}(y|x), \quad \text{[5 pts]}$$

with the Bayes risk being the conditional entropy of Y given X:

$$L(f, P) = \mathbb{E}_{(X,Y)\sim P}[-\log p_{Y|X}(Y|X)] = H_P(Y|X). \quad \text{[5 pts]}$$

Hint: You might find the following fact useful: For any two distributions P and Q on $\mathcal{X}$, the KL (Kullback-Leibler) divergence $D(P\|Q)$ between P and Q, defined as

$$D(P\|Q) \triangleq \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)},$$

is always non negative, i.e.,

$$D(P\|Q) \geq 0$$

where the inequality holds with equality iff $P = Q$.

PROBLEM 2 (BIAS-VARIANCE TRADEOFF; 30 PTS). The bias-variance tradeoff is a fundamental tradeoff in statistics and machine learning. To understand this tradeoff, in this problem we show that under the square loss, the error in both the parametric estimation and regression learning problems can decomposed as the sum of a (squared) bias and a variance term.

(a) First consider the parametric estimation problem in statistical decision theory. Assume an outcome space $\mathcal{X}$ and a class of probability distributions $\{P_\theta : \theta \in \Theta\}$ on the space $\mathcal{X}$. Assume we observe an outcome X generated by P_θ for some θ that is unknown to us. Based on this observation X, we want to estimate $g(\theta)$ for an arbitrary function g on Θ using some decision rule δ. Show that for any decision rule δ, the estimation error $\mathbb{E}_\theta[(g(\theta) - \delta(X))^2]$ under the square loss can be decomposed as:

$$\mathbb{E}_\theta[(g(\theta) - \delta(X))^2] = (\text{Bias}_\theta(\delta))^2 + \text{Var}_\theta(\delta), \qquad \text{[10 pts]}$$

where

$$\text{Bias}_\theta(\delta) \triangleq g(\theta) - \mathbb{E}_\theta[\delta(X)]$$

is the bias of estimator δ and

$$\text{Var}_\theta(\delta) \triangleq \mathbb{E}_\theta[(\mathbb{E}_\theta[\delta(X)] - \delta(X))^2]$$

is the variance of the estimator δ.

If an estimator δ satisfies $\mathbb{E}_\theta[\delta(X)] = g(\theta)$, then we say δ is an unbiased estimator. Based on the above bias-variance decomposition, explain why the performance of an unbiased estimator is determined by its variance. [5 pts]

(b) Now consider the regression problem in machine learning. Assume $Y = f(X) + W$, where X is the input and $W \sim \mathcal{N}(0, N)$ is a Gaussian noise that is independent of X. We want to find, based on an i.i.d. generated sequence of n training examples $Z^n = \{(X_i, Y_i)\}_{i=1}^n$, a predictor $h_{Z^n} \in \mathcal{H}$ that achieves small squared prediction error

$$\mathbb{E}_{Z^n,X,Y}[(Y - h_{Z^n}(X))^2].$$

Show that the following decomposition of the error holds:

$$\mathbb{E}_{Z^n,X,Y}[(Y - h_{Z^n}(X))^2] = \text{Bias}^2 + \text{Var} + \text{Bayes Risk}, \qquad \text{[10 pts]}$$

where

$$\text{Bias}^2 \triangleq \mathbb{E}_X[(f(X) - \mathbb{E}_{Z^n}[h_{Z^n}(X)])^2]$$

is the bias of the learner and

$$\text{Var} \triangleq \mathbb{E}_{Z^n,X}[(\mathbb{E}_{Z^n}[h_{Z^n}(X)] - h_{Z^n}(X))^2]$$

is the variance of the learner, and

$$\text{Bayes Risk} = N$$

is the irreducible Bayes risk.

Can you make a plot to demonstrate how the bias, variance, and overall error vary with the complexity of the model $\mathcal{H}$? [5 pts]

PROBLEM 3 (NO-FREE-LUNCH THEOREM; 20 PTS). Recall the following form of the NFL theorem we have shown in Lecture 4:

"Let A be any learning algorithm for the task of binary classification with respect to 0-1 loss over a domain $\mathcal{X}$. Let n be any number small than $|\mathcal{X}|/2$, representing the training set size. Then there exists a distribution P over $\mathcal{X} \times \{0, 1\}$ such that

$$\mathbb{E}[L(A(Z^n), P)] \geq 1/4."$$

In this problem, we show that the above statement implies the version of the NFL theorem stated in Theorem 1.1 of Lecture 4.

(a) First recall the Markov inequality, which says that for any nonnegative random variable X and $a \geq 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Use the Markov inequality to prove the following reverse Markov inequality: For any random variable Y that takes value on $[0, 1]$ and $a \in (0, 1)$,

$$\mathbb{P}(Y \geq a) \geq \frac{\mathbb{E}[Y] - a}{1 - a}. \qquad \text{[10 pts]}$$

(b) Use the above reverse Markov inequality to show that if $\mathbb{E}[L(A(Z^n), P)] \geq 1/4$, then $P^n(L(A(Z^n), P) \geq 1/8) \geq 1/7$. [10 pts]

PROBLEM 4 (VC DIMENSION; 20 PTS). In Lecture 5, we have shown that for a finite hypothesis class $\mathcal{H}$, VC-d$(\mathcal{H}) \leq \log |\mathcal{H}|$. Now consider the domain set $\mathcal{X} = [0, 1]$.

(a) Find an example of $\mathcal{H}$ on $\mathcal{X}$ so that $\mathcal{H}$ is infinite while VC-d$(\mathcal{H}) = 1$. [10 pts]

(b) Find an example of $\mathcal{H}$ on $\mathcal{X}$ so that VC-d$(\mathcal{H}) = \log |\mathcal{H}| = 2$. [10 pts]