Previously when describing "how does supervised learning work", we have

- A training data $\{\mathbf{x}_i, y_i\}_{i=1}^n$
- A regression/classification function f
- A loss function $L(y_i, f(\mathbf{x}_i))$ that evaluates the performance of f on one sample $(\mathbf{x}, y)$
- training error: averaged loss over the n training samples

$$\frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i))$$

- test error: averaged loss over N future test samples

$$\frac{1}{N}\sum_{j=1}^{N} L(y_j^*, f(\mathbf{x}_j^*))$$

- Suppose $\{x_j^*, y_j^*\}_{j=1}^N$ is a set of test data with test error given by

$$\text{TestErr}[f] = \frac{1}{N} \sum_{j=1}^{N} \left[y_j^* - f(x^*)\right]^2.$$

  Naturally, we would like to have a very large test set. If $N \to \infty$, the average above is equivalent to $\mathbb{E}(Y^* - f(X^*))^2$ where $(X^*, Y^*) \sim$ some distribution $P(x, y)$. Although this distribution P is often unknown in practice, we can still mathematically formulate the test error as an expectation.

- It's reasonable to assume the training data as iid samples from the same unknown distribution. If the training and test data are governed by totally different random processes, then learning is not possible.

- There are some learning algorithms that try to extract knowledge from one domain and then adapt it to other domains, but even those algorithms assume that something are shared across different domains.
- As a simple illustration, let's assume the training and test data follow the same distribution P.
- Next let's take a glimpse of statistical decision theory for supervised learning.

Statistical Decision Theory

- Assume $(X, Y) \sim$ some distribution $P(x, y)$.
- We have a loss function to evaluate the prediction accuracy of f,

  for regression, e.g., $L(y, f(x)) = (y - f(x))^2$,

  for classification [a], e.g., $L(y, f(x)) = 0$ if $y = f(x)$ and 1 otherwise.
- The final criterion is the averaged loss (a.k.a., Risk)

$$R[f] = \mathbb{E}_{X,Y} L(Y, f(X)).$$

  The optimal f is given by $f^* = \arg\min R[f]$ and the optimal risk is denoted by $R^* = R[f^*] = \min_f R[f]$ (often called the Bayes risk).

[a] Consider only hard decision rules here.

Assume the joint distribution P is known. What's the optimal f^*?

$$\begin{aligned} R[f] &= \mathbb{E}_{X,Y} L(Y, f(X)) \\ &= \mathbb{E}_X \big[\mathbb{E}_{Y|X} L(Y, f(X))\big]. \quad (*) \end{aligned}$$

Given $X = x$, the inside expectation $\mathbb{E}_{Y|X=x}$ is over Y only.[a] You can view $(*)$ as

$$\mathbb{E}_X \big[\mathbb{E}_{Y|X} L(Y, f(X))\big] = \int_{\mathcal{X}} \Big[\int_{\mathcal{Y}} L(y, f(x)) p(y|x) dy\Big] p(x) dx$$

where $p(x)$ is the marginal distribution function of X and $p(y|x)$ is the conditional distribution function of Y given $X = x$.

[a] Law of iterated expectations. For a quick review on conditional probabilities, check my 410 notes [Note_0901] on this page.

$$
\begin{aligned}
R[f] &= \mathbb{E}_X \left[\mathbb{E}_{Y|X} L(Y, f(X)) \right] \\
&= \int_{\mathcal{X}} \left[\int_{\mathcal{Y}} L(y, f(x)) p(y|x) dy \right] p(x) dx
\end{aligned}
$$

The problem of finding an optimal function f that minimizes $R[f]$

$\Longrightarrow$ reduced to a series of sub-problems: find the optimal value of $f(x)$ to minimize the inside expectation for each given x.

$$
f^*(x) = \arg\min_a \mathbb{E}_{Y|X=x} L(Y, a). \quad \text{😇}
$$

We can do this for every x, and then the resulting f^* (of course, it may not be continuous) minimizes $R[f]$.

😇 turns out to be not difficult to solve: Y is just a one-dimensional (in regression) or discrete (in classification) random variable. Of course the solution would depend on the loss function L.

- For regression with squared loss,

$$f^*(x) = \arg\min_a \mathbb{E}_{Y|X=x}(Y-a)^2 = \mathbb{E}[Y \mid X = x]$$

  What if we use absolute loss, $L(y, f(\mathbf{x})) = |y - f(x)|$?

- For classification with $0/1$ loss,

$$f^*(x) = 1, \text{ if } P(Y = 1 \mid X = x) > 0.5.$$

  This is known as the Bayes classifier (due to the connection with the Bayes' theorem). This is why we call the corresponding risk, $R[f^*]$, the Bayes risk.

- In practice, however, we don't know P (so cannot calculate f^*), but we are provided with a set of random samples $(x_i, y_i)_{i=1}^n$ from P. And we usually restrict our regression/classification function f to some family $\mathcal{F}$.

If we have all the data (i.e., know the data generating process P), we have

- the risk of a function f: $R[f] = \mathbb{E}_{X,Y} L(Y, f(X))$
- the optimal function f^* with the optimal risk:

$$f^* = \arg\min_f R[f], \quad R^* = R[f^*].$$

- the optimal function and optimal risk wrt the function space $\mathcal{F}$:

$$f^*_{\mathcal{F}} = \arg\min_{f \in \mathcal{F}} R[f], \quad R^*_{\mathcal{F}} = R[f^*_{\mathcal{F}}]$$

Given a set of training data of size n, all we have are

- the empirical risk of a function f: $\hat{R}_n[f] = \frac{1}{n} \sum_{i=1}^n L(y_i, f(\mathbf{x}_i))$
- the empirical optimal function wrt the function space $\mathcal{F}$:

$$\hat{f}_{n,\mathcal{F}} = \arg\min_{f \in \mathcal{F}} \hat{R}_n[f].$$

A natural question: how far is $R[\hat{f}_{n,\mathcal{F}}]$ from the ideal performance R^*? [a]

$$\begin{aligned} R[\hat{f}_{n,\mathcal{F}}] - R^* &= R[\hat{f}_{n,\mathcal{F}}] - R[f^*_{\mathcal{F}}] + R[f^*_{\mathcal{F}}] - R^* \\ &= \text{Variance} + \text{Bias} \end{aligned}$$

The first term can be decomposed as

$$\begin{aligned} & R[\hat{f}_{n,\mathcal{F}}] - \hat{R}_n[\hat{f}_{n,\mathcal{F}}] + \hat{R}_n[\hat{f}_{n,\mathcal{F}}] - \hat{R}_n[f^*_{\mathcal{F}}] + \hat{R}_n[f^*_{\mathcal{F}}] - R[f^*_{\mathcal{F}}] \\ \leq\ & R[\hat{f}_{n,\mathcal{F}}] - \hat{R}_n[\hat{f}_{n,\mathcal{F}}] + \hat{R}_n[f^*_{\mathcal{F}}] - R[f^*_{\mathcal{F}}] \\ \leq\ & 2 \max_{f \in \mathcal{F}} |R[f] - \hat{R}_n[f]| \end{aligned}$$

Note that $R[f]$ is the mean of $\hat{R}_n[f]$, so the first term represents the variance wrt the function space $\mathcal{F}$.

[a] Usually we do not care whether f^* is close to $\hat{f}_{n,\mathcal{F}}$. Actually f^* may not even be unique. All we care is the closeness in terms of their risk.