Statistical Learning C2

1.
Y =f(X)+ε

f is some fixed but unknown function of X1,…,Xp, ε is a random error term
statistical learning refers to a set of approaches for estimating f.

2.1.1 为什么估计 f?

(1) prediction

$\widehat{\mathrm{Y}} = \hat{\mathrm{f}}(\mathrm{X})$ where $\hat{\mathrm{f}}$ represents our estimate for f , and $\widehat{\mathrm{Y}}$ represents the resulting prediction for Y. $\hat{\mathrm{f}}$ is often treated as a black box.

$\hat{Y}$的准确性和 reducible error，irreducible error 有关。Reducible error 可以改进 f 函数。但是因为 Y 也是关于 ε 的函数，不能靠 X 预测，所以这部分 error 是 irreducible error。

$$\begin{aligned} E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\ &= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}} \quad , \end{aligned}$$

(2) inference
什么预测变量与结果有关？结果和每个预测变量是什么关系？预测变量和结果可以只用线性方程总结，还是更复杂？

2.1.2 怎么估计 f

(1) Parametric Method(two-step model-based approach)
先假设模型，第二部 use training data to fit or train the model。可以选择 flexible models 提高准确，
(2) Non-parametric Methods
不做函数假设，直接训练模型。缺点：需要很多数据。

2.2.1 评估 model 准确性

对于 regression，用 mean squared error(MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

MSE is computed using the training data. But we do not really care how well the method works on the training data, we are interested in the accuracy of predictions to unseen test data.

当一个方法产生小的 training MSE but 大的 test MSE，就是 overfitting。但是无论 overfitting 是否发生，training MSE 都要比 test MSE 小。

2.2.2 Bias-Variance trade-off

Variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set. Variance 是不同的训练数据集训练出的模型输出值之间的差异。

Bias 是用所有可能的训练数据集训练出的所有模型的输出的平均值与真实模型的输出值之间的差异。

要找 both variance and squared bias are low 的 method

2.2.3 classification

Training error rate: proportion of mistakes that are made we apply our estimate $\hat{f}$ to the training observations:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i).$$

$\hat{y}_i$ 是估计的结果，如果 yi 不等于$\hat{y}_i$，I 等于 1.yi=$\hat{y}_i$i，I 等于 0

(1) The Bayes Classifier

叠加 indicator 作为条件，使得 Pr(Y=j | X=x0)Pr(Y=j | X=x0) 这个概率最大。例如，在 binary classification 问题上，如果 Pr(Y=1 | X=x0)>0.5 则预测 class one，否则预测 class two。

(2) K-Nearest Neighbors