Elements of statistical learning Ch. 7 notes

James Chuang

April 27, 2017

Contents

My notes on The Elements of Statistical Learning Ch. 7 on Model Assessment and Selection

7.1 Introduction

- ***generalization*** performance of a learning method: its prediction capability on independent test data
  - guides the choice of learning method or model
  - a measure of the quality of the ultimately chosen model

7.2 Bias, Variance, and Model Complexity

- Consider the case of a quantitative or interval scale response with:
  - target variable Y
  - vectors of inputs X
  - prediction model $\hat{f}(X)$ estimated from a training set τ
  - a loss function for measuring errors between Y and $\hat{f}(X)$: $L\left(Y, \hat{f}(X)\right)$
    - typical choices:

$$L\left(Y, \hat{f}(X)\right) = \begin{cases} \left(Y - \hat{f}(X)\right)^2 & \text{squared error} \\ \left|Y - \hat{f}(X)\right| & \text{absolute error} \end{cases}$$

  - ***test error***, aka ***generalization error***: the prediction error over an independent test sample

$$\text{Err}_{\mathcal{T}} = \text{E}\left[L\left(Y, \hat{f}(X)\right) \mid \mathcal{T}\right]$$

    - X and Y drawn randomly from their joint distribution (population)
    - here the training set τ is fixed, and the test error is for this specific training set
  - ***expected prediction error***, aka ***expected test error***:

$$\begin{aligned} \text{Err} &= \text{E}\left[L\left(Y, \hat{f}(X)\right)\right] \\ &= \text{E}\left[\text{Err}_{\mathcal{T}}\right] \end{aligned}$$

    - expectation averages over everything that is random, including randomness in the training set that produced $\hat{f}$
    - goal: estimate Err_{τ}. However, Err is more amenable to statistical analysis
  - ***training error***: the average loss over the training sample:

$$\overline{\text{err}} = \frac{1}{N}\sum_{i=1}^{n} L\left(y_i, \hat{f}(x_i)\right)$$

- want to know the expected test error of the estimated model $\hat{f}$
  - a more complex model uses the training data more and is able to adapt to more complicated underlying structures
    - hence, there is a decrease in bias but an increase in variance
    - some intermediate model complexity will give minimum expected test error
- training error is **not** a good estimate of the test error
  - training error consistently decreases with model complexity, and eventually drops to zero with enough complexity
    - a model with zero training error is overfit to the training data and will typically generalize poorly

- Similarly, consider a qualitative or categorical response G taking one of K values in a set $\mathcal{G}$, labeled for convenience as $1, 2, \ldots, K$
  - typically, we model the probabilities $p_k(X) = \text{Pr}\left(G = k \mid X\right)$ (or some monotone transformations $f_k(X)$)
  - then, classify by $\hat{G}(X) = \arg\max_k \hat{p}_k(X)$
    - in some cases (e.g. 1-nearest neighbor classification), $\hat{G}(X)$ is produced directly
  - typical loss functions:

$$L\left(G, \hat{G}(X)\right) = I\left(G \neq \hat{G}(X)\right) \qquad \text{(0-1 loss)}$$

$$\begin{aligned} L\left(G, \hat{p}(X)\right) &= -2\sum_{k=1}^{K} I\left(G = k\right)\log \hat{p}_k(X) \\ &= -2\log \hat{p}_G(X) \qquad (-2 \times \text{log-likelihood, aka the } \textbf{deviance}) \end{aligned}$$

  - again, ***test error*** and the expected misclassification error:

$$\text{Err}_{\mathcal{T}} = \text{E}\left[L\left(G, \hat{G}(X)\right) \mid \mathcal{T}\right], \qquad \text{Err} = \text{E}\left[\text{Err}_{\mathcal{T}}\right]$$

  - training error is the sample analogue, e.g.:

$$\overline{\text{err}} = -\frac{2}{N}\sum_{i=1}^{N}\log \hat{p}_{g_i}(x_i) \quad \text{sample log-likelihood for the model}$$

  - the log-likelihood can be used as a loss function for general response densities, e.g. Poisson, gamma, exponential, log-normal and others
  - if $\text{Pr}_{\theta(X)}(Y)$ is the density of Y, indexed by a parameter $\theta(X)$ that depends on the predictor X, then

$$L\left(Y, \theta(X)\right) = -2 \cdot \log \text{Pr}_{\theta(X)}\left(Y\right)$$

    - "-2" in the definition makes the log-likelihood loss for the Gaussian distribution match squared-error loss

- notation for the rest of the chapter:
  - Y and $f(X)$ represent all of the above situations, since the focus is mainly on the quantitative response (squared-error loss) setting
  - typically, a model will have tuning parameter(s) (hyperparameters) α, so predictions can be written $\hat{f}_\alpha(x)$
    - tuning parameter varies the complexity of the model
    - want to find the value of α that minimizes error, i.e. produces the minimum of the average test error curve
    - for brevity, the dependence of $\hat{f}(x)$ on α is often suppressed

- two separate goals:
  - ***model selection***: estimating the performance of different models in order to choose the best one
  - ***model assessment***: having chosen a final model, estimating its prediction error (generalization error) on new data
  - in a data-rich situation, the best approach for both problems is to randomly divide dataset into training, validation, and test sets
    - training set is used to fit the models
    - validation set used to estimate prediction error for model selection
    - test set used for assessment of the generalization error of the final chosen model
      - the test set should **only** be brought out at the end of the data analysis
  - methods of this chapter either approximate the validation step analytically (AIC, BIC, MDL, SRM) or by efficient sample re-use (cross-validation and the bootstrap)
    - these methods are used in model selection and also provide an estimate of the test error of the final chosen model

7.3 The Bias-Variance Decomposition

- Assumptions:
  - $Y = f(X) + \epsilon$
    - $\mathsf{E}[\epsilon] = 0$
    - $\mathsf{Var}(\epsilon) = \sigma_\epsilon^2$
- we can derive an expression for the expected prediction error of a regression fit $\hat{f}(X)$ at an input point $X = x_0$, using squared error loss:

$$
\begin{aligned}
& \mathsf{Err}(x_0) \\
&= \mathsf{E}\left[\left(Y - \hat{f}(x_0)\right)^2 \mid X = x_0\right] \\
&= \mathsf{E}\left[\left(f(X) + \epsilon - \hat{f}(x_0)\right)^2 \mid X = x_0\right] && Y = f(X) + \epsilon \\
&= \mathsf{E}\left[\left(\left(f(X) - \hat{f}(x_0)\right) + \epsilon\right)^2 \mid X = x_0\right] \\
&= \mathsf{E}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2 + 2\epsilon\left(f(x_0) - \hat{f}(x_0)\right) + \epsilon^2\right] \\
&= \mathsf{E}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right] + 2\mathsf{E}\left[\epsilon\left(f(x_0) - \hat{f}(x_0)\right)\right] + \mathsf{E}\left[\epsilon^2\right] && \text{expectations are linear} \\
&= \mathsf{E}\left[\left(f(x_0) - \mathsf{E}\hat{f}(x_0) + \mathsf{E}\hat{f}(x_0) + \hat{f}(x_0)\right)^2\right] + 2\mathsf{E}\left[\epsilon\right]\mathsf{E}\left[\left(f(x_0) - \hat{f}(x_0)\right)\right] + \mathsf{E}\left[(\epsilon - \mathsf{E}\left[\epsilon\right])^2\right] && \mathsf{E}[\epsilon] = 0 \\
&= \mathsf{E}\left[\left(f(x_0) - \mathsf{E}\hat{f}(x_0)\right)^2 + 2\left(f(x_0) - \mathsf{E}\hat{f}(x_0)\right)\left(\mathsf{E}\hat{f}(x_0) - \hat{f}(x_0)\right) + \left(\mathsf{E}\hat{f}(x_0) - \hat{f}(x_0)\right)^2\right] + 0 + \sigma_\epsilon^2 && \mathsf{E}[\epsilon] = 0 \text{ again} \\
&= \mathsf{E}\left[\left(f(x_0) - \mathsf{E}\hat{f}(x_0)\right)^2\right] + 2\mathsf{E}\left[\left(f(x_0) - \mathsf{E}\hat{f}(x_0)\right)\left(\mathsf{E}\hat{f}(x_0) - \hat{f}(x_0)\right)\right] + \mathsf{E}\left[\left(\mathsf{E}\hat{f}(x_0) - \hat{f}(x_0)\right)^2\right] + \sigma_\epsilon^2 && \text{expectations are linear} \\
&= \mathsf{Bias}^2\left(\hat{f}(x_0)\right) + \mathsf{Var}\left(\hat{f}(x_0)\right) + \sigma_\epsilon^2 && \mathsf{E}\left[\mathsf{E}\hat{f}(x_0) - \hat{f}(x_0) \mid X \right. \\
&= \mathsf{Bias}^2 + \mathsf{Variance} + \mathsf{Irreducible\ Error}
\end{aligned}
$$

- ***squared bias***: the amount by which the average of the estimate differs from the true mean
- ***variance***: the expected squared deviation of $\hat{f}(x_0)$ around its mean
- ***irreducible error***: variance of the target around its true mean $f(x_0)$
  - cannot be avoided no matter how well we estimate $f(x_0)$, unless $\sigma_\epsilon^2 = 0$
- typically, the more complex a model $\hat{f}$, the lower the (squared) bias but the higher the variance
- for k-nearest-neighbor regression fit, the error has the simple form:

$$
\begin{aligned}
\mathsf{Err}(x_0) &= \mathsf{E}\left[\left(Y - \hat{f}_k(x_0)\right)^2 \mid X = x_0\right] \\
&= \sigma_\epsilon^2 + \left[f(x_0) - \frac{1}{k}\sum_{\ell=1}^{k} f\left(x_{(\ell)}\right)\right]^2 + \frac{\sigma_\epsilon^2}{k}
\end{aligned}
$$

- here we assume that the training inputs x_i are fixed, and the randomness arises from the y_i
  - the number of neighbors k is inversely related to the model complexity:
    - for small k, the estimate $\hat{f}_k(x)$ can potentially adapt itself better to the underlying $f(x)$
    - as k is increased, the bias – the squared difference between $f(x_0)$ and the average of $f(x)$ at the k-nearest neighbors – will typically increase, while the variance decreases
- for a linear model fit $\hat{f}_p(x) = x^T\hat{\beta}$, where the parameter p-vector β is fit by least squares:

$$
\begin{aligned}
\mathrm{Err}(x_0) &= \mathsf{E}\left[\left(Y-\hat{f}_p\left(x_0\right)\right)^2 \mid X=x_0\right] \\
&= \sigma_\epsilon^2+\left[f\left(x_0\right)-\mathsf{E}\hat{f}_p(x_0)\right]^2+\|\mathbf{h}(x_0)\|^2\,\sigma_\epsilon^2
\end{aligned}
$$

$$
\begin{aligned}
&\mathbf{h}(x_0)=\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}x_0, \quad \text{the } N\text{-vector of linear weights producing the fit:} \\
&\hat{f}_p(x_0)=x_0^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}, \quad \text{hence,} \\
&\mathrm{Var}\left[\hat{f}_p\left(x_0\right)\right]=\left\|\mathbf{h}\left(x_0\right)\right\|^2\sigma_\epsilon^2
\end{aligned}
$$

- while this variance changes with x_0, its average (with x_0 taken to be each of the sample values x_i) is $\left(\frac{p}{N}\right)\sigma_\epsilon^2$, hence the ***in-sample error*** is:

$$
\frac{1}{N}\sum_{i=1}^{n}\mathrm{Err}(x_i)=\sigma_\epsilon^2+\frac{1}{N}\sum_{i=1}^{N}\left[f\left(x_i\right)-\mathsf{E}\hat{f}\left(x_i\right)\right]^2+\frac{p}{N}\sigma_\epsilon^2
$$

- here, model complexity is directly related to the number of parameters p