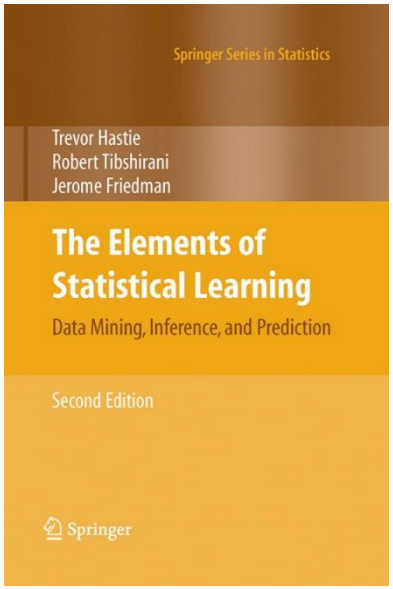
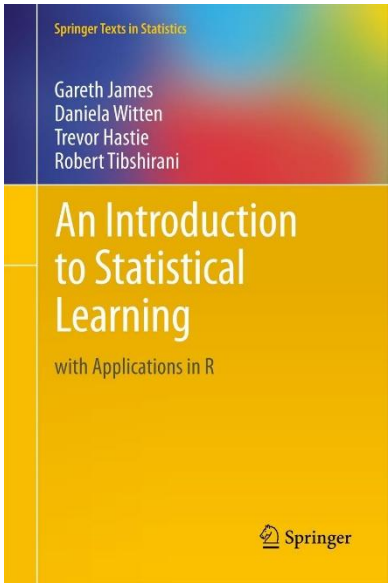
The Elements of Statistical Learning

- **http://www-stat.stanford.edu/~tibs/ElemStatLearn/download.html**
- **http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf**





Elements of Statistical Learning

Printing 10 with corrections

- Contents

Elements of Statistical Learning

There is no true interpretation of anything; interpretation is a vehicle in the service of human comprehension. The value of interpretation is in enabling others to fruitfully think about an idea.

–Andreas Buja

- **Supervised**
  - *Least Squares*
  - *Nearest Neighbor*
- **Unsupervised**

Elements of Statistical Learning

- **Basics**
  - *Supervised Learning*
    - **Outcome/Output (usually quantitative)**
    - **feature set (measurements)**
    - **training data (feature measurements plus known outcome)**
  - *Unsupervised Learning*
    - **Outcome is unavailable or unknown**

  - *Example problems available at http://www-stat.stanford.edu/ElemStatLearn*
    - **Is it spam?**
    - **prostate cancer**
    - **digit recognition**
    - **microarray**

Elements of Statistical Learning

- **Basics/Terminology**
  - *variable types*
    - **quantitative**
    - **qualitative (AKA categorical, discrete, factors)**
      - *values in a finite set, G* = {Virginica, Setosa and Versicolor}
      - *classes – 10 digits, G* = {0, 1, . . . , 9}
    - **ordered categorical**
      - *G = { small, medium, large }*
      - *not metric*
      - *typically represented as codes, e.g., {0,1} or {-1,1} (AKA targets)*

        most useful encoding - dummy variables, e.g. vector of K bits to for K-level qualitative measurement or more efficient version
  - *Symbols/terminology*
    - **Input variable (feature),** X
    - **Output variable**
      - *Quantitative,* Y
      - *Qualitative,* G
    - **variables can be matrices (bold) or vectors. All vectors are column vectors**
    - **hat (e.g.,** $\hat{Y}$ **)variables are estimates**

Elements of Statistical Learning

- **Linear Models (least squares) (2.3.1)**
  - *input* $X^T = (X_1, X_2 \ \ldots, X_p)$
  - *output* $Y^T = (Y_1, Y_2 \ \ldots, Y_p)$

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \, \hat{\beta}_j$$

  - *intercept is the bias,* $\hat{\beta}_0$
    - **if included in** X**, this can be written as an inner product (dot product), assumed from now on**

$$\big((x_0, x_1, \ldots, x_n), (y_0, y_1, \ldots, y_n)\big) = (x_0, x_1, \ldots, x_n) \cdot (y_0, y_1, \ldots, y_n) = x_0 y_0 + x_1 y_1 + \ldots + x_n y_n$$

$$\hat{Y} = f(X) = X^T \hat{\beta}$$

Elements of Statistical Learning

- **Linear Models (least squares)**
  - *Minimize residual sum of squares, RSS*

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} \left(y_i - x_i^T \beta\right)^2$$

$$\mathrm{RSS}(\beta) = (y - \boldsymbol{X}\beta)^T (y - \boldsymbol{X}\beta)$$

  - *taking the derivative and setting it to zero (unless the $\boldsymbol{X}^T\boldsymbol{X}$ is singular*)*

$$\hat{\beta} = \frac{\boldsymbol{X}^T y}{\boldsymbol{X}^T \boldsymbol{X}}$$

$$\hat{y}_i = \hat{y}(x_i) = x_i^T \hat{\beta}$$

* *determinant is zero, has no inverse*

Elements of Statistical Learning

- **Linear Models (least squares) for classification**
  - *output G = { Blue, Orange } = {0,1}*

$$G = \begin{cases} \text{Orange} & \text{if} \quad \hat{Y} > 0.5 \\ \text{Blue} & \text{if} \quad \hat{Y} \leq 0.5 \end{cases}$$

  - *if data comes from two bivariate Gaussian distributions solution is close to optimum (linear solution is best, Chapter 4)*
  - *if dta comes from Gaussian mixture, not close to optimal*

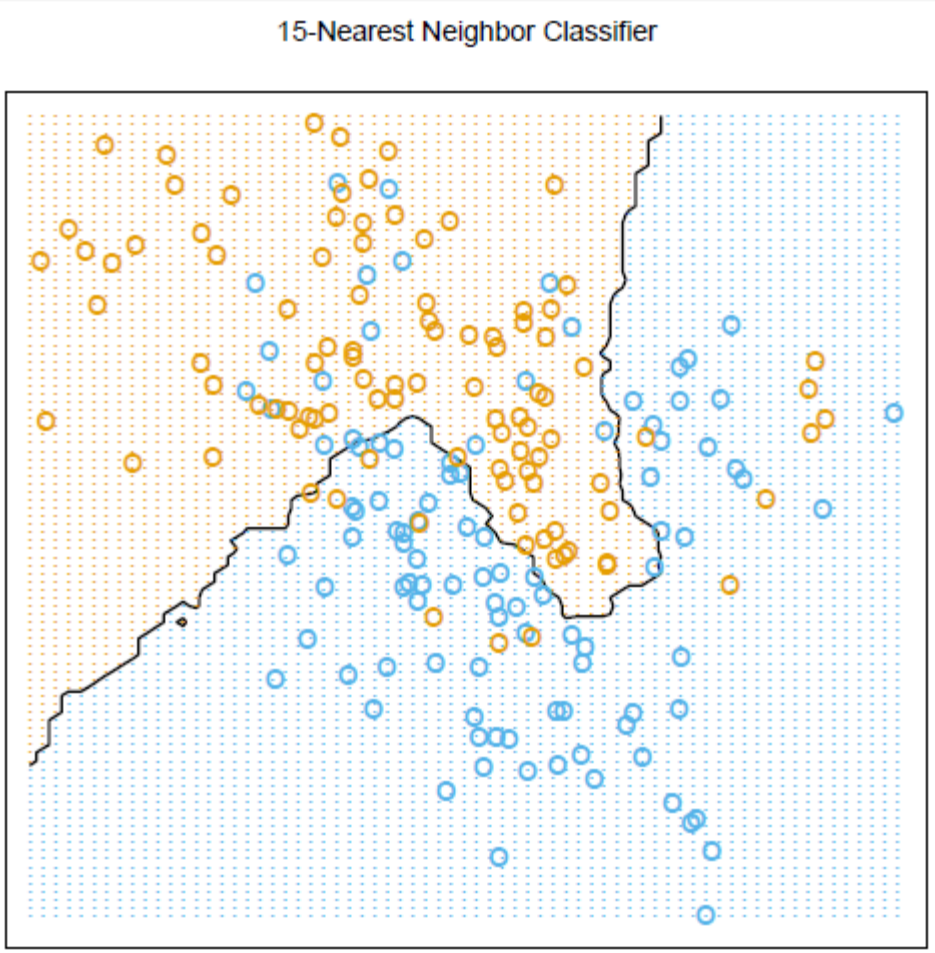
Elements of Statistical Learning

- **Nearest neighbor methods (2.3.2)**
  - *output is assigned based on the nearest observations in the training set, $\mathcal{T}$*

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

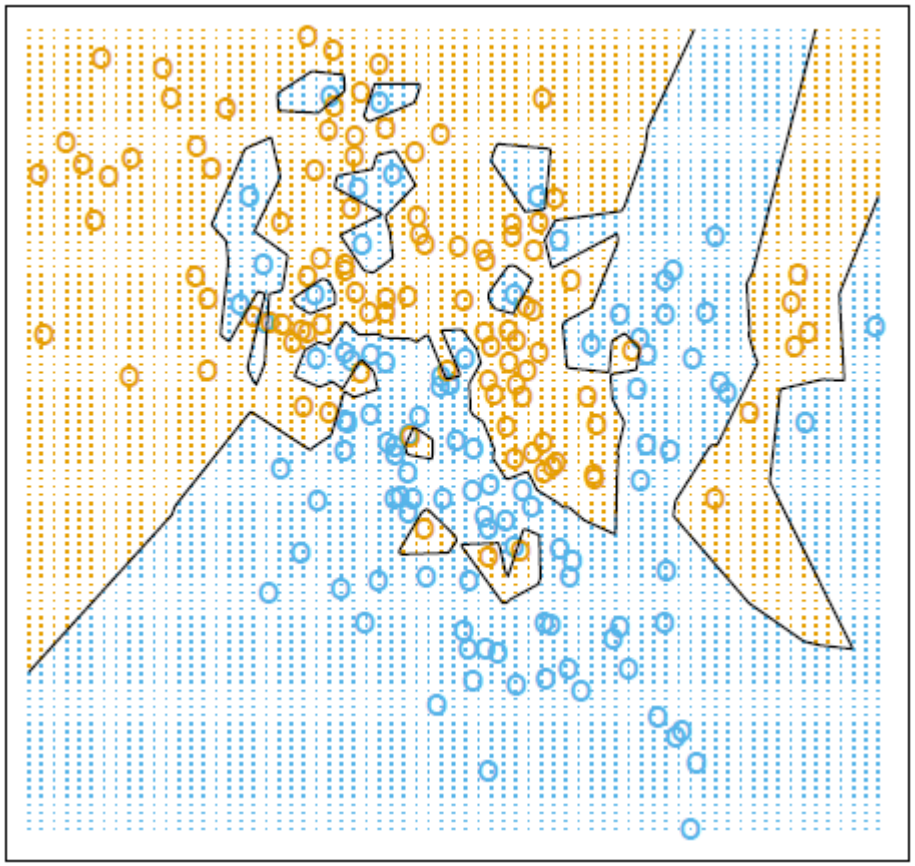
where $N_k(x)$ is the neighborhood of the k closest points to x_i

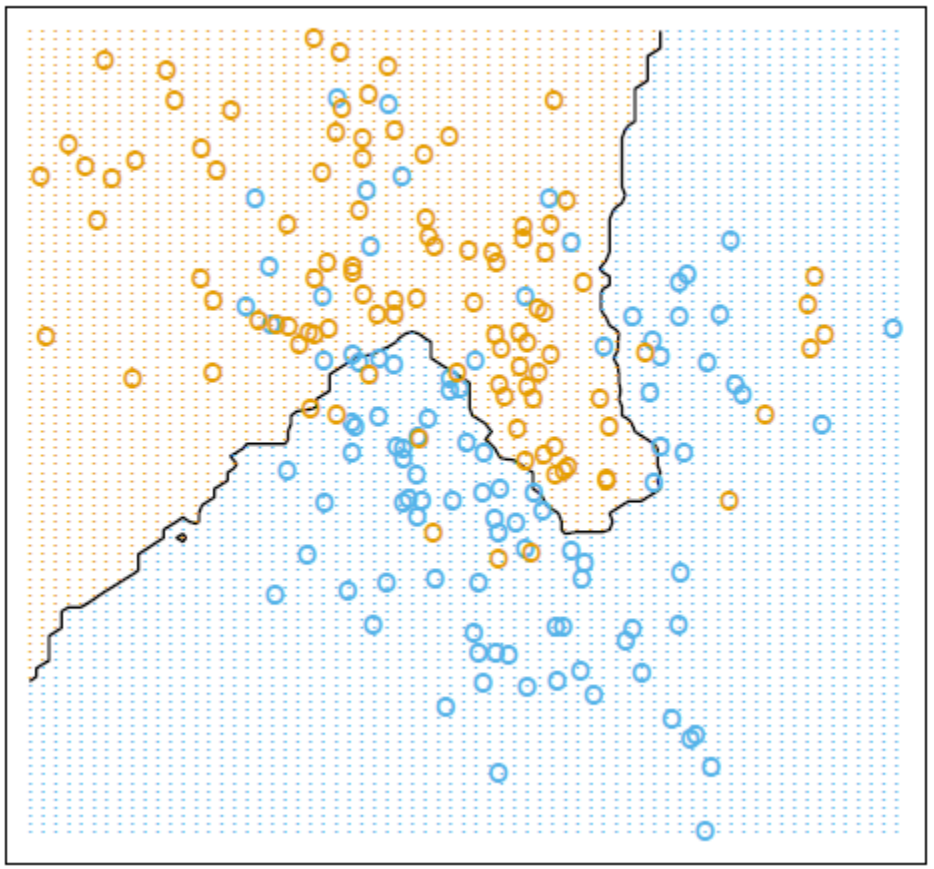
  - *Apparently less misclassification, but the effective number of parameters ($^N/_k$) is different*
  - *error is zero when k=1*



Elements of Statistical Learning

- **Nearest neighbor methods**



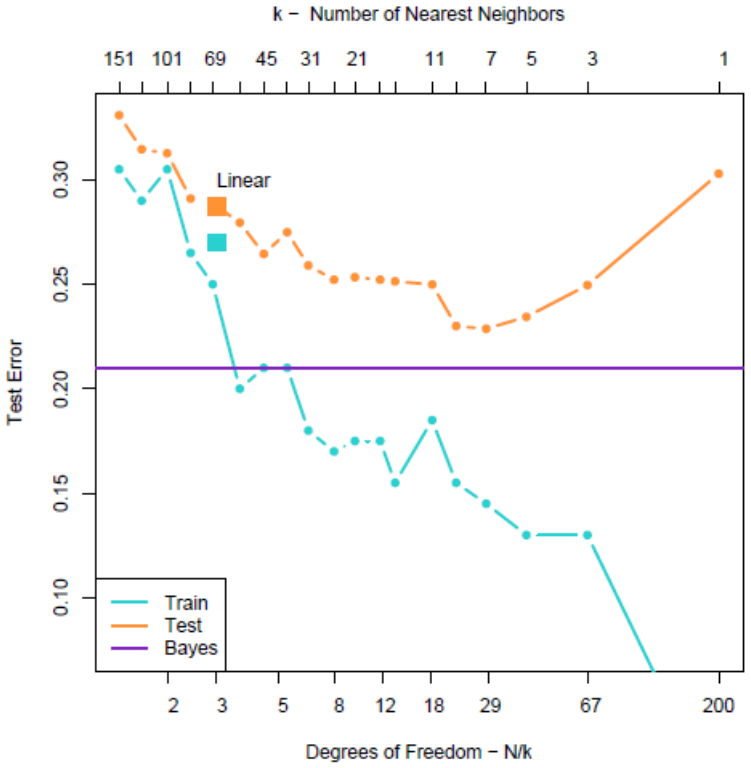


Elements of Statistical Learning

- **Choosing k**
  - *Simulated data*
  - *10 Blue means generated from N((1, 0)^T , I)*
  - *10 Orange means generated from N((0,1)^T , I)*
  - *For each class, generate 100 observations*
    - **pick a mean randomly, P=0.1**
    - **generate a Gaussian distributed random observation with *N(m_k, I/5)***
  - *training data = 200 points*
  - *test data = 10,000 points*
  - *Bayes error rate (purple)*

$$p = \sum_{C_i \neq C_{max}} \int_{x \in H_i} P(x|C_i)p(C_i)dx$$

where C_i is a class and H is the area classified as C_i



Elements of Statistical Learning

- 2.4 Statistical Decision Theory
- 2.5 Local Methods in High Dimensions
- 2.6 Statistical Models, Supervised Learning and Function Approximation
- 2.7 Structured Regression Models
- 2.8 Classes of Restricted Estimators
- 2.9 Model Selection and the Bias–Variance Tradeoff

Elements of Statistical Learning

- **2.4 Statistical Decision Theory**
  - *X* a real valued input vector $X \in \mathbb{R}^p$
  - *Y* a real valued output variable $Y \in \mathbb{R}$
  - Joint probability $P(X, Y)$
  - A function for predicting *Y*, $f(X) = \hat{Y}$

  - The prediction is optimized by minimizing a *loss function,* $L\big(Y, f(X)\big)$
  - Commonly *squared error*, $L\big(Y, f(X)\big) = \big(Y - f(X)\big)^2$
    Or *absolute error* $L\big(Y, f(X)\big) = Y - f(X)$

Elements of Statistical Learning

- **Expected Prediction Error (EPE), eq 2.9**

$$EPE(f) = E\left(\left(Y - f(X)\right)^2\right)$$
$$= \int |y - f(x)|^2 \, P(x, y) dx dy$$

- **Eq 2.10 rewritten slightly**
- **By Bayes rule** $P(X, Y) = P(Y|X)P(X)$, substituting into the integral gives

$$= \int |y - f(x)|^2 \, P(x|y) P(x) dx dy$$

$$= \int P(x) \left(\int |y - f(x)|^2 \, P(x|y) dy \right) dx$$

$$= E_X \, E_{Y|X} \left(|Y - f(X)|^2 | X \right) \quad \text{eq 2.11}$$

Elements of Statistical Learning

- **EPE is minimized (for squared error) when**

$$f(x) = E(Y|X = x)$$

  - *Regression equation 2.13, the best prediction of Y at a point X=x is the conditional mean*

  - *Nearest neighbor methods*
    - **Nearest neighbor methods are a direct implementation where E(Y|X) is estimated as the average of all the values in the neighborhood of x**

$$\hat{f}(x) = \text{Ave}\ (\ y_i|\ x_i \in N_k(x)\ )$$

    - **As k becomes large, there are many points in the neighborhood and the average becomes stable**
    - **Why is this not a perfect estimator?**
      - *Sample size is often small*
      - *Neighborhood grows with the size of the feature vector p, neighborhood is a poor surrogate for conditioning*

Elements of Statistical Learning

- **Linear Regression**

$$f(x) = x^T\beta$$

Substitute into equation and take the derivative to find where the Expected Prediction Error is minimized

$$\frac{\partial\, EPE}{\partial\beta} = -2\int (y - x^T\beta)x\,\Pr(dx, dy) = 0$$

$$= E(yx) - E(xx^T\beta) = 0$$

$$\beta = E(xx^T)^{-1}E(yx)$$

Which is eq 2.16, which amounts. The least squares equation (2.6) replace the expectations with averages

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

Elements of Statistical Learning

- **Comparison of least squares and k-neighbors**
  - *Least squares - f(x) is well approximated by a globally linear function*
    - **Stable but biased**
  - *K-nearest neighbors – f(x) is well approximated by a locally constant function*
    - **more flexible**
    - **Less stable but less biased**

  - *Other loss functions can be used, e.g. L_1*
  - *Categorical functions with similar derivation give the Bayes classifier*

Elements of Statistical Learning

- **2.5 Local Methods in High Dimensions / Bias - Variance Decomposition**
  - *In higher dimensions, the fraction of the range of each features that must be used to cover a fixed fractional volume becomes very large – finding local neighborhoods becomes difficult*
  - *Bias - Variance Decomposition*
  - $\mathrm{MSE}(x_0) = E_\tau[f(x_0) - \hat{y}_0]^2$

$$
\begin{aligned}
&= E_\tau[\hat{y}_0 - E_\tau(\hat{y}_0) + E_\tau(\hat{y}_0) - f(x_0)]^2 \\
&= E_\tau\left[\left(\hat{y}_0 - E_\tau(\hat{y}_0)\right)^2 + 2\left(\hat{y}_0 - E_\tau(\hat{y}_0)\right)\left(E_\tau(\hat{y}_0) - f(x_0)\right) + \left(E_\tau(\hat{y}_0) - f(x_0)\right)^2\right] \\
&= E_\tau\left[\left(\hat{y}_0 - E_\tau(\hat{y}_0)\right)^2\right] + \left(E_\tau(\hat{y}_0) - f(x_0)\right)^2 \\
&= \text{variance} + \text{bias}^2
\end{aligned}
$$

  - *"By imposing some heavy restrictions on the class of models being fitted, we have avoided the curse of dimensionality"*

Elements of Statistical Learning

- **2.6 Statistical Models, Supervised Learning and Function Approximation**
  - *Find a function, $\hat{f}(x)$, that approximates the true function $f(x)$ that relates inputs and outputs*
  - *Regression function results from squared loss error model*
    - **Nearest neighbor methods are direct estimates of conditional expectation**
    - **In dimension is high, nearest neighbors may not be close (large errors)**
    - **Special structure can be used to reduce both variance and bias**
  - *For a real statistical model, $Y = f(X) + \varepsilon$*
    - **If error is independent of X, P(Y|X) depends on X only through the conditional mean**
    - **But error , variance, and bias do not need to be independent**

Elements of Statistical Learning

- **2.6.2 Supervised Learning**
  - *For this book = function approximation*
    - **If samples size is large and dense, all solutions tend to limiting conditional expectation**
    - **when N is finite, eligible solutions must be restricted**
  - *Maximum likelihood estimation is an alternative to least squares*
  - *Notation: parameters generally referred to as θ*
  - *Classes of model, each has smoothing parameters (regularizers)*
    - **Bayesian methods/roughness penalty**
    - **Kernel methods/local regression**
    - **Basis Functions/Dictionary methods**

Elements of Statistical Learning

- **Model choice**
  - *Can't use RSS, because interpolating methods will always give good fit*
  - *although RSS is small, bias and variance may be large*
  - *for k-nearest-neighbor regression*

$$\mathrm{EPE}_k(x_0) = \sigma^2 + \left[f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f(x_l) \right]^2 + \frac{\sigma^2}{k}$$

irreducible error

bias (MSE)

variance average of neighborhood

  - *bias increases as k increases*
    variance decreases as k increases
  - *Choose model to minimize test error (not training error)*

Elements of Statistical Learning

- **Bias-variance tradeoff**

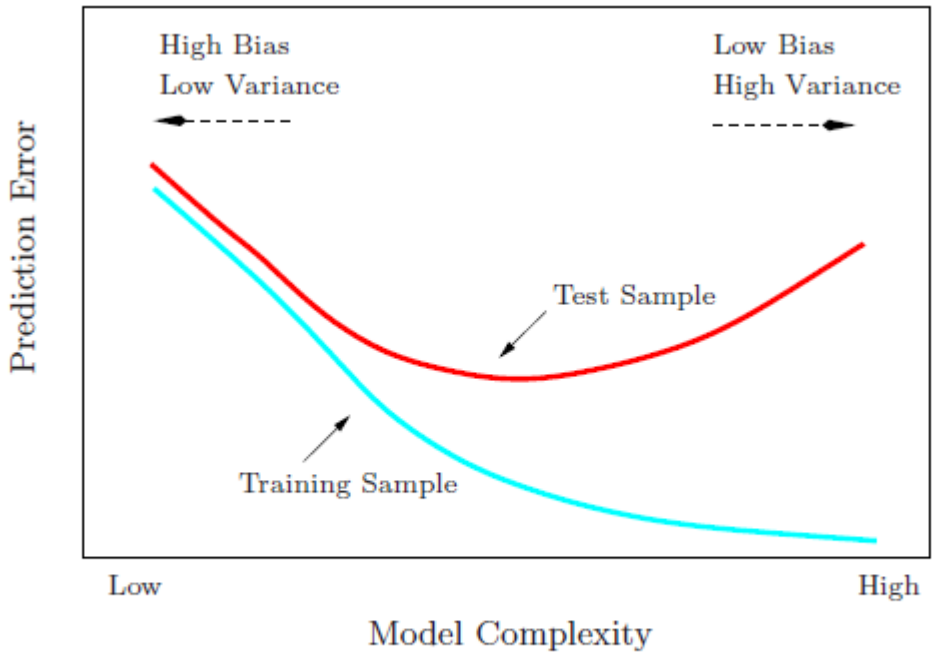


FIGURE 2.11. *Test and training error as a function of model complexity.*

Elements of Statistical Learning

- **ex 2.1**
- **K classes, each with a target t_k, each t_k is all zeroes except position *k*, for instance $t_1 = [0, 1, 0, \dots]$**
  for an observation $\hat{y}$ with $|\hat{y}| = 1$ (not stated, all positive, i.e. probability)
  distance to a target is $d = \|\hat{y} - t_k\|$
  this is a little difficult to expand because of the square root in the norm, so you can recognize that

$$\begin{aligned}\text{argmin}_k d = \text{argmin}_k d^2 &= \text{argmin}_k \|\hat{y} - t_k\|^2 \\ &= \text{argmin}_k \sum_{l=1}^{K} \left(\hat{y}_l - t_{k,l}\right)^2 \\ &= \text{argmin}_k \sum_{l=1} \left(\hat{y}_l^2 - 2\hat{y}_l t_{k,l} + t_{k,l}^2\right)\end{aligned}$$

- **since the sums of $\hat{y}_l^2$ and $t_{k,l}^2$ are both constants, they don't affect the min and can be ignored. also note $2\hat{y}_l t_{k,l} = 0$ except when $l = k$, so**

$$\begin{aligned}&= \text{argmin}_k(-2\hat{y}_k) \\ &= \text{argmax}_k\, \hat{y}_k\end{aligned}$$