Chapter 1

- Stat learning - vast set of tools for understanding data
- Supervised learning - involves building a model for predicting an output based on 1+ inputs
- unsupervised learning - there are inputs but no supervising output → we try to learn patterns from the data
- regression problem - continuous / quantative output
- classification problem - categorical or qualative output
- clustering problem - find groups based on observed characteristics of inputs

★ PICK METHODS based on response!

Chapter 5

- Resampling - involve repeatedly drawing samples from training set

→ model assessment: evaluating a model's performance

→ model selection: selecting proper level of flexibility

★ Bootstrap - accuracy of parameters (non parametric)

★ CV = estimate test error associated or level of flexibility

- Validation set approach: randomly split into 2 sets
  - → different results depend on different splits
  - → validation estimate of test error highly variable
  - → only trained w/ subset = over estimates test error
- Leave-One-Out-CV (LOOCV): single observation used for validation
  - → unbiased estimated of test error
  - → $CV_n = \frac{1}{n}\sum_{i=1}^{n} MSE$ ↗ tends not to overestimate test error
  - → advantage: less bias than validation, always yeilds same results
  - → disadvantage: time consuming, $CV_n = \frac{1}{n}\sum\left(\frac{y_i-\hat{y}_i}{1-h_i}\right)$ least squares
- k fold = randomly divide into k-folds, fit k-1 folds
  $CV_k = \frac{1}{k}\sum MSE$, LOOCV = special case $k=n$, adv: computations

Validation Set ← bias — k-fold — variance → LOOCV

Chapter 3: Regression, supervised learning

SLR = $Y \approx B_0 + B_1 X \rightarrow \hat{y} = \hat{B}_0 + \hat{B}_1 X$ (least squares line)

$\hat{B}_0 = \bar{y} - \hat{B}_1\bar{x}$

$\bar{y} = \frac{1}{n}\sum y_i$

$\bar{x} = \frac{1}{n} x_i$

→ least squares criterion: $e_i = y_i - \hat{y}_i$ (measuring closeness)

→ RSS: $e_1^2 + e_2^2 + \ldots + e_n^2 = (y_1 - \hat{B}_0 - \hat{B}_1 x) + \ldots + (y_n - B_0 - B_1 \hat{x}_n)$

⇒ least square approach, minimize RSS: $\hat{B}_1 = \sum_{i=1}^{n}\frac{(x_i-\bar{x})(y_i-\bar{y})}{(x_1-\bar{x})^2}$

→ Standard Error:

$SE(\hat{B}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right]$, $Var(\hat{u}) = SE(\hat{u})^2 = \frac{\sigma^2}{n}$ - Standard Deviation

$SE(\hat{B}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_0-\bar{x})^2}$ | $\sigma^2 = var(\epsilon)$, σ, $RSE = \sqrt{\frac{RSS}{n-2}}$

Confidence Interval: 95% CI = range of values such 95% probabilities that the range will contain true unknown value

→ in LR: $\hat{B} \pm 2\cdot SE(\hat{B})$ | $H_0: B_1 = 0$ | t-stat: $t = \frac{\hat{B}_1 - 0}{SE(\hat{B})}$

Hypothesis Testing | $H_a = B_1 \neq 0$ | p-value: probability of $x \geq |t|$

→ small p value - reject null hypothesis

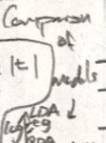
Model Accuracy: model fit → $RSE = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{1}{n-2}\sum(y_i-\hat{y}_i)^2}$

RSE (lack of fit): deviation from regression line.

R^2: independent from scale of Y, measured in proportion (variance explained)

$R^2 = \frac{TSS - RSS}{TSS} = \frac{ESS}{TSS}$ | $TSS = \sum(y_i - \bar{y})^2$ ⇒ 0: did not explain, 1 = has explained

explained



MLR: why not run seperate SLR: ① single predictions? ② predictors ignore others! | attr regress..

$\hat{Y} = \hat{B}_0 + \hat{B}_1 X_1 + \hat{B}_2 X_2 + \ldots + \hat{B}_p X_p$: minimize $RSS = \sum(y_i - \hat{y}_i)^2$

$H_0: B_1 = B_2 = B_p = 0$, H_a: at least 1 B_i is non zero | pvalue of Fstat important!

F stat: $\frac{TSS - RSS/P}{RSS - (n-p-1)}$ } $H_0 \Rightarrow f \approx 1$, $H_a \Rightarrow f > 1$ > large n 1+, small n 1++

⇒ R^2 always increases when p increases, $RSE = \sqrt{\frac{RSS}{n-p-1}}$

models can have a higher RSE, if decrease in RSS is small relative to increase in p with more vars

predictions: uncertainty

① ↑ estimate! coefficent in accuracy → confidence intervals

② model bias → assumpting shape of function

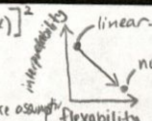
③ irreducible error → prediction intervals Y from ŷ ↓ larger than CI.

Chapter 2

$Y = f(X) + \epsilon$ — random error term / independent of X, mean (0)

$f(X)$: systematic information (goal!)

$E(Y-\hat{Y})^2 = E[f(x) + \epsilon - \hat{f}(x)]^2 = \underbrace{[f(x) - \hat{f}(x)]^2}_{\text{reducible error}} + \underbrace{var(\epsilon)}_{\text{irreducible}}$



linear → better for influence

non-linear → better for prediction

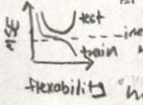
increasing flexibility ↗ increases complexity → estimate can be poor! → overfit! learning the noise

- Parametric Methods: 2 step model based approach ① make assumption ② estimate parameters
  ⇒ reduces problem of estimating f down to estimating set of parameters.
  advantage: simplifying problem, disadvantage = will not match the true function ⇒
- Non-Parametric Method: do not make explicit assumptions about functional forms, instead they choose to estimate f that gets as close to data points as possible.
  advantage: potential to accurately fit wider range of fs. disadvantage = large # of n needed to obtain estimate
  lower level of smoothness = better model fit: overfitting

Model Accuracy:

→ regression setting - quantify the extent to which predicted response is close to the response

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



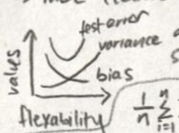
Bias-Variance Trade-off:

$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + (Bias(\hat{f}(x_0)))^2 + Var(\epsilon)$ ↙ ★ test MSE cannot be lower than

→ lowest Test MSE = low variance + low bias (ideal)

- variance = amount by which $\hat{f}$ changes if we use diffrent training data
  high variance → small changes in training data lead to large in f, higher = more flex
- bias = error that is introduced by approx. real world problem w/ simpler model, lower = more flex

⇒ more flexability = higher variance, lower bias



as flexability increases, initially variance increases but bias decreases faster so MSE falls but after a while bias has little impact & variance increases faster, MSE increases!

$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$ ↗ I=0 ✓ ↘ I=1 × } fraction of incorrect classification ⇒ test error rate: $Ave(I(y_0 \neq \hat{y}_0))$ is smallest!

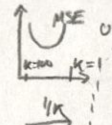
→ classification setting ⇒ assign observation to the most likely class, given its predictor values:

Bayes Classifier: test error minimized when: $Pr(Y=j \mid X=0)$ is largest / lowest possible test error rate:

→ Bayes Decision Boundary = probability is exactly 50% / Bayes error rate = $1 - E(\max_j Pr(Y=j \mid X=x_0))$ = irreducible error

ideal scenario because conditional probability ↑

KNN: $Pr(Y=j \mid X=x_0) = \frac{1}{K}\sum_{i \in N} I(y_i = j)$
- ① identifies neighbors
- ② estimates conditional probability for class j
- ③ applies Bayes rule
- ④ classifies w/ highest prob.



over fit as k is small

CV in classification

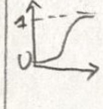
$CV_{(n)} = \frac{1}{n}\sum(I(y_i \neq \hat{y}_i))$

Bootstrap: unknown, SE, w/ a given estimate

$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2}$

Chapter 4: Classification (supervised) ⇒ why not LR: ① no natural order ② binary setting better but can lead to unrealistic p(x)

Logistic Regression: models the probabilities that Y belongs to a class rather than class directly.



$p(X) = \frac{e^{B_0 + B_1 X}}{1 + e^{B_0 + B_1 X}}$ (logistic function) ⇒ $\left[\frac{p(x)}{1-p(x)}\right]$ = odds = $e^{B_0 + B_1 X}$ B_1 gives → change in log odds.

0 = low prob, ∞ = high pr

⇒ $\log\left(\frac{p(x)}{1-p(x)}\right)$ = log odds = $B_0 + B_1 X$

p(x) depends on current value

maximum likelihood function:

we seek $B_0 + B_1$ so $\hat{p}(x_i)$ for classes correspond as close as possible to observed!

$l(B_0, B_1) = \prod_{i: y_i = 1} p(x_i) \prod_{i: y_i = 0}(1 - p(x_i))$ ⇒ accuracy of coeffecien — z-stat = $z(B_1) = \frac{B_1}{SE(B_1)}$

→ large z, small p, reject H_0

⇒ confounding: correlation between predictors.

$H_0: \frac{e^{B_0}}{1+B^0}$

LDA: multiple classes ⇒ less direct approach: models distributions of predictors in each class, Bayes Theorem to flip into $Pr(Y=k \mid X=x)$

→ why not log reg: classes not well separated, n is small, more than 2 response classes.

→ Prior Probability = π_k (randomly chosen n comes from class k

→ Density function = $f_k(x) = Pr(X=x \mid Y=k)$ ⇒ high prob if n from k has X = x.

→ Posterior Prob = $P_k(x) = Pr(Y=k \mid X=x) = \frac{\pi_k f_k(x)}{\sum \pi_l f_l(x)} = \frac{\pi_k (f_k(x))}{\pi_k f_k(x) \pi_l f_l(x)}$ - assume normal

QDA

no common $\Sigma_1 \neq \Sigma_2 = \Sigma_3$

use when training set large

LDA for p=1

assume $f_k(x)$ is normal: $f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(\frac{-1}{2\sigma_k^2}(x-u_k)^2\right)$ — mean, variance

$\sigma_1^2 = \sigma_2^2 = \sigma_k^2$ (shared variance)

⇒ assign n to largest:

$\delta_k(x) = x\cdot\frac{u_k}{\sigma^2} - \frac{u_k^2}{2\sigma^2} + \log(\pi_k)$

when $K=2$, $\pi_1 = \pi_2$: $x = \frac{u_1^2 + u_2}{2}$ { u_k = mean Ⓚ, $\sigma^2 = \frac{\sum(x_i - u_k)^2}{n-k}$ ① Ⓚ, $\pi_k = n_k/n$

growing model → σ^2 high

LDA for p>1

$u_k = 1/n_k \sum x_i$ (p·k)

$\sigma^2 = \frac{1}{n-k}\sum\sum(x_i - u_k)^2$ $\left(\frac{p(p-1)}{2}\right)$

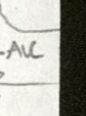
Σ Ⓚ

$\delta_k(x) = x^T\Sigma^{-1} - \frac{1}{2}u_k^T\Sigma^{-1}u_k + \log(\pi_k)$

	P	
Y	TN	FN
Y	FP	TP

$\frac{TN + TP}{n}$ = accuracy

$\frac{FP + FN}{n}$ = error rate



Qualatative Predictors

dummy = levels - 1

$y_i = B_0 + B_1 X_1 + B_2 X_2 + \epsilon$

↓

$B_0 + B_1 + \epsilon$ (class A)

$B_0 + B_2 + \epsilon$ (class B)

$B_0 + \epsilon$ (class C)

Extensions

- additive: predictors = independent

⇒ add interaction, hierarchal principle include main effect

- linear assumptions

→ transform X

⇒ more R^2 (better fit)

→ might overfit.

$Y = B_0 + B_1 X + B_2 X^2 + \epsilon$

Potential Problems

① non-linearity of R-P relationships

② correlation error terms

③ non-constant variance error

④ outliers: extreme Y

⑤ high leverage: extreme X

⑥ collinearity - dependence