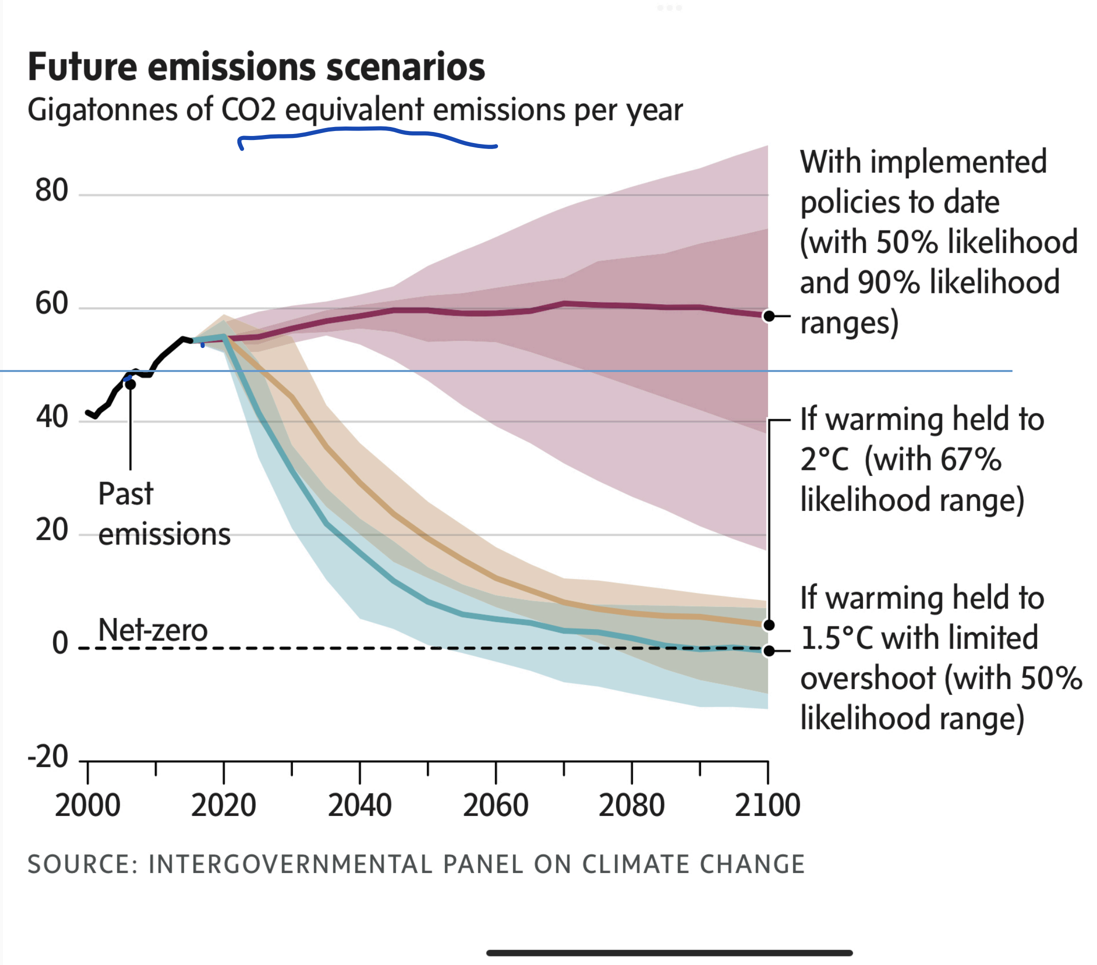
Mathematical Statistics II

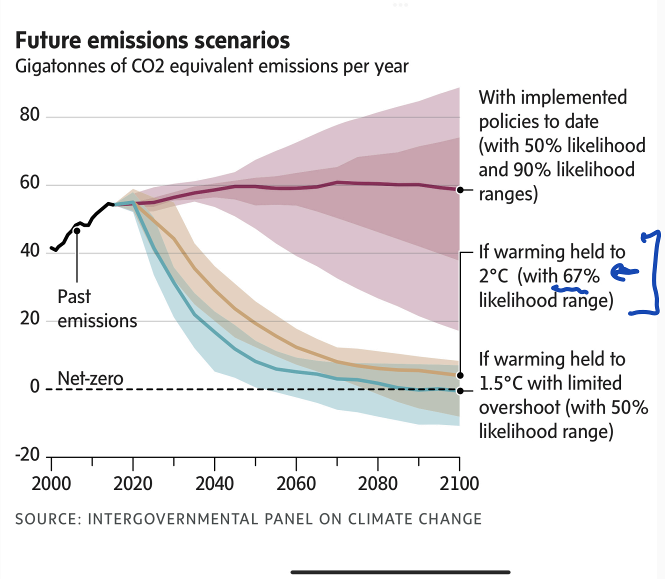
STA2212H S LEC0101

Week 11

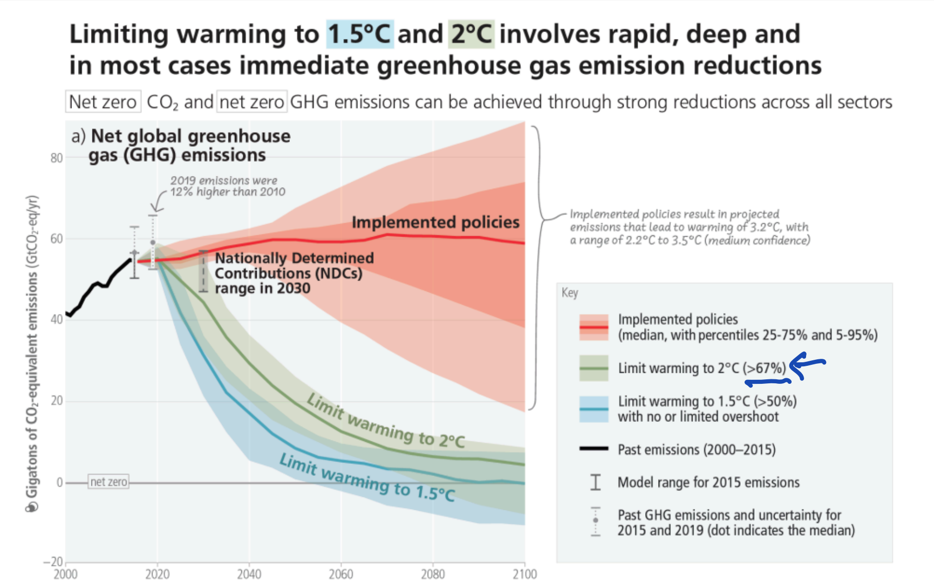
March 28 2023



IPCC Report



Future emissions scenarios
Gigatonnes of CO2 equivalent emissions per year
80
60
40
20
0
-20
2000
2020
2040
2060
2080
2100
Past emissions
Net-zero
With implemented policies to date (with 50% likelihood and 90% likelihood ranges)
If warming held to 2°C (with 67% likelihood range)
If warming held to 1.5°C with limited overshoot (with 50% likelihood range)
SOURCE: INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE



Limiting warming to 1.5°C and 2°C involves rapid, deep and in most cases immediate greenhouse gas emission reductions
Net zero CO₂ and net zero GHG emissions can be achieved through strong reductions across all sectors
a) Net global greenhouse gas (GHG) emissions
Gigatons of CO₂-equivalent emissions (GtCO₂-eq/yr)
2019 emissions were 12% higher than 2010
Implemented policies
Nationally Determined Contributions (NDCs) range in 2030
Limit warming to 2°C
Limit warming to 1.5°C
net zero
Implemented policies result in projected emissions that lead to warming of 3.2°C, with a range of 2.2°C to 3.5°C (medium confidence)
Key
Implemented policies (median, with percentiles 25-75% and 5-95%)
Limit warming to 2°C (>67%)
Limit warming to 1.5°C (>50%) with no or limited overshoot
Past emissions (2000–2015)
Model range for 2015 emissions
Past GHG emissions and uncertainty for 2015 and 2019 (dot indicates the median)
80
60
40
20
0
−20
2000
2020
2040
2060
2080
2100

Today

[t,

1. Recap
2. Directed acyclic graphs
3. Aspects of classification

\beamertemplatenavigationsymbols{empty}

\\~\\

Upcoming

- April 4 10.00 – 13.00 Math Stat II Project presentations

please submit slides (pdf) by April 3

- April 11 9.30 – 12.30 Hydro Room 9016
  Informal discussion of large language models

Recap

- observational studies can provide some evidence towards causality
- but care must be taken re confounding variables

  Simpson's "paradox"

  trt   response

- if all confounding variables are adjusted for, we have stronger evidence of the causal effect of a treatment on outcome
- this requires an assumption of "no unmeasured confounding"
- Bradford-Hill guidelines for strengthening support for causality

  in the absence of randomized treatment assignment

- one popular approach to causality is through the notion of counterfactuals

$E(C_1 - C_0)$

AoS

- the causal treatment effect is $\theta = E(Y(1) - Y(0))$; the difference in outcome for an individual with $X = 1$ compared to her outcome with $X = 0$

C_0, C_1

- also called the causal risk difference

risk ratio   odds ratio

- since both outcomes cannot be observed, we must assume that in our data the units are "similar enough" that we can average over the treated and control to estimate θ

$C_1, C_0 \perp\!\!\!\perp X \mid Z$

- $\alpha = \mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0)$ estimated by $\bar{Y}_1 - \bar{Y}_0$

association

$\alpha \neq \theta$

unless

$(C_0\ C_1) \perp X \mid Z$

- causal risk difference $\theta = Y(1) - Y(0)$
- causal risk difference as a function of an additional covariate Z

$$\theta(z) = \mathrm{E}(Y(1) \mid Z = z) - \mathrm{E}(Y(0) \mid Z = z) = \mathrm{E}(C_1 \mid Z = z) - \mathrm{E}(C_0 \mid Z = z)$$

r.v. w a dist^n

Thm 16.6: no unmeasured confounding

continuous exposure X

- causal regression function

$\boxed{C(x) = \beta x \ldots}$

$E(Y|x)$ $\downarrow$ in x

$$\theta(x) = \mathrm{E}\{C(x)\}$$

- association function

$\boxed{C_i(x) = \beta_i x}$

$$r(x) = \mathrm{E}(Y \mid X = x)$$

$\theta(x) \neq r(x)$ unless

- Thm 16.4: If X assigned at random $\theta(x) = r(x)$

? $C(x) \perp X \mid Z$?

- graphs can be useful for clarifying dependence relations among random variables

SM Markov random fields

- a Directed Acyclic Graph has random variables on the vertices and edges joining random variables

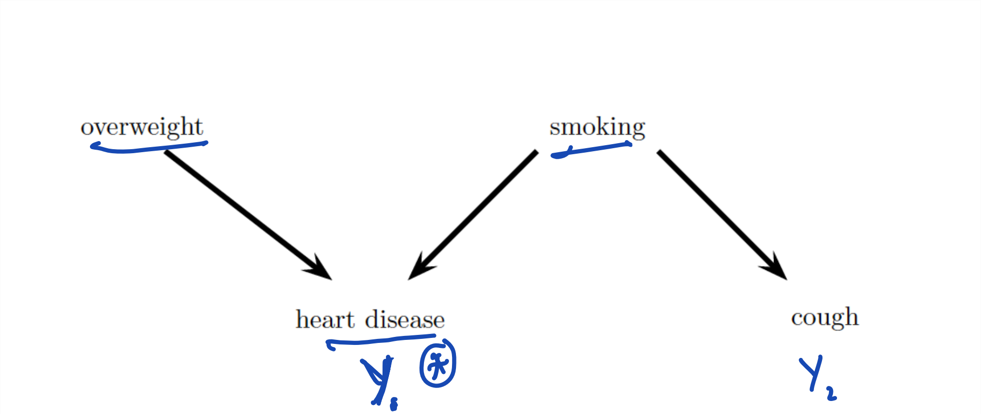
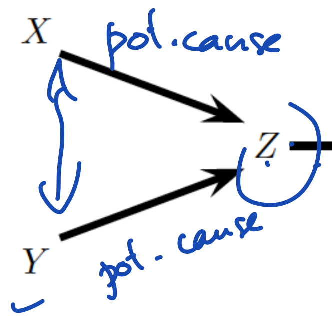


FIGURE 17.2. DAG for Example 17.4.

model says



$f(x, y, z, w)$

$= f(w \mid z) f(z \mid x, y)$

$f(x) f(y)$

FIGURE 17.3. Another DAG.

encodes a ~~too~~ Markov property of cond'l indep

Directed graphs and conditional independence

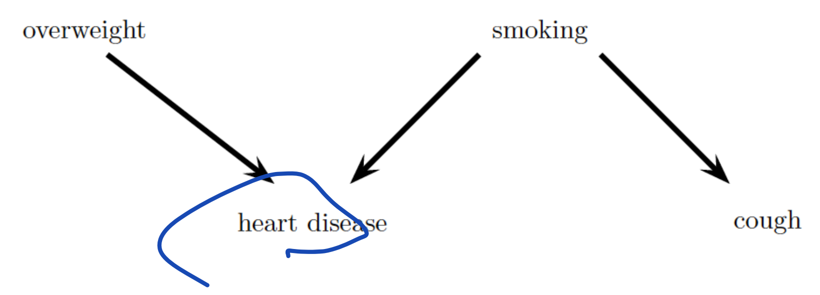


FIGURE 17.2. DAG for Example 17.4.

17.4 Example. Figure 17.2 shows a DAG with four variables. The probability function for this example factors as

$$\begin{aligned} f(\text{overweight}, \text{smoking}, \text{heart disease}, \text{cough}) & \\ = \; & f(\text{overweight}) \times f(\text{smoking}) \\ \times \; & f(\text{heart disease} \mid \text{overweight}, \text{smoking}) \\ \times \; & f(\text{cough} \mid \text{smoking}). \quad \blacksquare \end{aligned}$$

x's

17.5 Example. For the DAG in Figure 17.3, $\mathbb{P} \in M(\mathcal{G})$ if and only if its probability function f has the form

$$f(x, y, z, w) = f(x) f(y) f(z \mid x, y) f(w \mid z). \quad \blacksquare$$

Directed graphs and causality

AoS 17.8

- variables at parent nodes are potential causes for responses at child nodes
- probability distribution on a DAG represents causality if and only if the probability distribution is Markov wrt the DAG

AoS 17.5; HR Ch 6

$f(y_t \mid y_{t-1}, \dots, y_1)$

$= f(y_t \mid y_{t-1})$

1-step Markov chain

- DAGs can be used to represent confounders

276 17. Directed Graphs and Conditional Independence

confounder

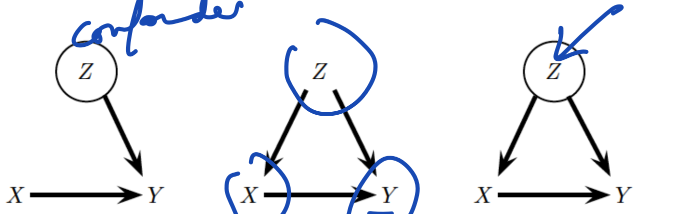


FIGURE 17.11. Randomized study; Observational study with measured confounders; Observational study with unmeasured confounders. The circled variables are unobserved.

Classification

$X_1, \dots, X_n \mid \text{iid}$ $f(x;\theta)$ $Z_1, \dots, Z_n$

- covariate space $\mathcal{X} \subset \mathbb{R}^d$; prediction space $\mathcal{Y} = \{0, 1\}$ — or $\mathcal{Y} = \{1, \dots, K\}$
- a classification rule

$$h : \mathcal{X} \to \mathcal{Y}$$

- data $(X_1, Y_1), \dots, (X_n, Y_n)$; $X_i \in \mathcal{X} \subset \mathbb{R}^d$, $Y_i \in \mathcal{Y}$ — supervised learning

350 22. Classification

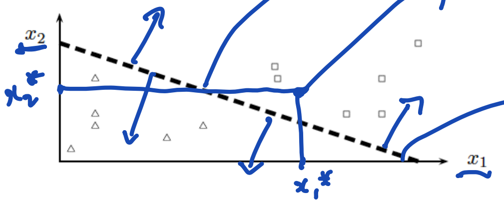


FIGURE 22.1. Two covariates and a linear decision boundary. $\triangle$ means $Y = 1$. $\square$ means $Y = 0$. These two groups are perfectly separated by the linear decision boundary; you probably won't see real data like this.

$\theta(h)$

$\mathcal{Y} = \{0, 1\}$

- loss function for classifier h:

$$L(h) = \mathrm{pr}\{h(X) \neq Y\}$$

$h(x) = \begin{cases} 0 \\ 1 \end{cases}$

- empirical error rate

training error rate

$$\hat{L}_n(h) = \frac{1}{n}\sum_{i=1}^{n} 1\{h(X_i) \neq Y_i\}$$

$Y_i = 0 \quad h(X_i) = 1$

$Y_i = 1 \quad h(X_i) = 0$

- special case $\mathcal{Y} = \{0, 1\}$ $\quad \hat{L}_n(h) = \frac{1}{n}\left[\#\{h(X_i) = 1,\ Y_i = 0\} + \#\{Y_i = 1, h(X_i) = 0\}\right]$

symmetric

- Bayes theorem

$$\mathrm{pr}(Y = 1 \mid X = x) = \frac{P(X = x \mid Y = 1)\,P(Y = 1)}{P(X = x \mid Y = 1)\,P(Y = 1) + \Pr(X = x \mid Y = 0)\,P(Y = 0)}$$

$f(x)\,dx = \Pr(X \in x, x + dx)$

$X \in \mathbb{R}^d \quad f(x)$

$$\Pr(Y=1 \mid X \in x, x+dx) = \frac{f_1(x)dx\,\pi_1}{f_1(x)dx\,\pi_1 + f_0(x)dx\,\pi_0}$$

$\pi_0 = 1-\pi_1$

- loss function for classifier h:

$$L(h) = \mathrm{pr}\{h(X) \neq Y\}$$

- empirical error rate

training error rate

$$\hat{L}_n(h) = \frac{1}{n}\sum_{i=1}^{n} 1\{h(X_i) \neq Y_i\}$$

- special case $\mathcal{Y} = \{0, 1\}$ $\quad \hat{L}_n(h) =$

symmetric

- Bayes theorem

$$\mathrm{pr}(Y = 1 \mid X = x) = \frac{f_1(x)\pi}{f_1(x)\pi + f_0(x)(1-\pi)}$$

Bayes classifier

- Bayes theorem

$$\mathrm{pr}(Y=1 \mid X=x) = \frac{f_1(x)\pi}{f_1(x)\pi + f_0(x)(1-\pi)} = r(x) \doteq \frac{1}{2}$$

$f_1(x)\pi \lessgtr \frac{1}{2}\{\quad\}$

- Bayes classifier

$$h^*(x) = \begin{cases} 1 & r(x) > 1/2 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 1 & \pi f_1(x) > (1-\pi) f_0(x) \\ 0 & \text{otherwise} \end{cases} = \frac{f_1(x)}{f_0(x)} > \frac{\pi}{1-\pi}$$

$\pi = \Pr(Y=1)$

- decision boundary

$$\mathcal{D} = \{x : \mathrm{pr}(Y=1 \mid X=x) = \mathrm{pr}(Y=0 \mid X=x)\}$$

- Thm 22.5: if h is another classification rule

$$L(h^*) \leq L(h)$$

Bayes risk

→ "posterior"

$$R(\tilde{\theta}, \theta) = \underset{x|\theta}{E}\{L(\tilde{\theta}(x), \theta)\}$$

$$L(\tilde{\theta}(x), \theta) = (\tilde{\theta}(x) - \theta)^2$$

$$E\{\qquad\} = E\{\qquad\} = \underline{\text{risk}} \; f =$$

Bayes risk $\int R(\tilde{\theta}, \theta)\, \pi(\theta)\, d\theta$

$$= \int\int L(\tilde{\theta}(\cancel{x}), \theta)\, \underline{f(\cancel{x}|\theta)\, dx \cdot \pi(\theta)}\, d\theta$$

$$= \int \underbrace{\int L(\tilde{\theta}(x), \theta)\, \pi(\theta|x)\, f_x(x)\, d\theta}\, dx$$

$$\min_{\tilde{\theta}} : \int L(\tilde{\theta}(x), \theta)\, \pi(\theta|x)\, d\theta = E_{\theta|x}\, L(\hat{\theta}(x), \theta)$$

- Bayes classifier:

$$h^*(x) = 1\{r(x) > 1/2\}, \quad r(x) = \frac{f_1(x)\pi}{f_1(x)\pi + f_0(x)(1-\pi)}$$

- multivariate normal

$$f_k(x) = \frac{1}{(2\pi)^{d/2}}|\Sigma_k|^{-1/2}\exp\{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\}$$

$k = 0, 1$

- $r(x) > 1/2 \iff$

$$\frac{\cancel{\frac{1}{(2\pi)^{d/2}}}|\Sigma_1|^{-1/2}\exp\{-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\}}{\ldots + \cancel{\frac{1}{(2\pi)^{d/2}}}|\Sigma_0|^{-1/2}\exp\{\ldots_{k=0}\}}$$

$\log > \log(\frac{1}{2})$

$x \in \mathbb{R}^d$

$x_j x_k \quad x_j^2 \quad x_k^2$

- $r(x) > 1/2 \iff$

QDA

$$(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) < (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) + \log(|\Sigma_0|/|\Sigma_1|) + 2\log\{\pi/(1-\pi)\}$$

$\pi_1 = \pi \quad \pi_0 = 1-\pi$

- estimates of π_k, μ_k, Σ_k:

MLE's for class 0

$$\hat\mu_0 = \frac{1}{n_0}\sum_{Y_i=0} X_i \qquad n_0 = \sum 1\{Y_i=0\}$$

$$\hat\Sigma_0 = \frac{1}{n_0}\sum_{Y_i=0}(X_i-\hat\mu_0)(X_i-\hat\mu_0)^T$$

$$\hat\pi_0 = \frac{1}{n}\sum 1\{Y_i=0\} \qquad \hat\pi_1 = \frac{1}{n}\sum 1\{Y_i=1\}$$

- $\Sigma_0 = \Sigma_1 = \Sigma$
- $\pi_0 = \pi_1$

$$r(x) > \tfrac{1}{2} \iff (x-\mu_1)^T\Sigma^{-1}(x-\mu_1) < (x-\mu_0)^T\Sigma^{-1}(x-\mu_0)$$

K-class prediction

$\hat{\mu}_1, \hat{\mu}_0, \hat{\Sigma}$

$x^T\Sigma_1^{-1}x + \Sigma^{-1}(x-\mu_1) + \mu_1^T\Sigma^{-1}\mu_1 < \Sigma^{-1}(x-\mu_0)\; x^T\Sigma_1^{-1}x + \mu_0^T\Sigma^{-1}\mu_0$

- If $\mathcal{Y} = \{1, \ldots, K\}$, the optimal classification rule is

$$
\begin{aligned}
h(x) &= \arg\max_k \operatorname{pr}(Y = k \mid X = x) \\
&= \arg\max_k \frac{\pi_k f_k(x)}{\Sigma_r \pi_r f_r(x)} \quad \text{— Bayes} \\
&= \arg\max_k \pi_k f_k(x)
\end{aligned}
$$

$r(x) > \frac{1}{2} \iff P(Y=1|x) > P(Y=0|x)$

$r(x) = \frac{1}{2}$ is a line in $\mathbb{R}^d$

- if $f_k(x)$ is Gaussian, then

$$h^*(x) = \arg\max_k \delta_k(x)$$

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log \pi_k$$

Linear Disc. Analysis

$\frac{d(d+1)}{2} + d$ params

Fisher's LDA

per class

- if $\mathcal{Y} = \{0, 1\}$ and $\Sigma_0 = \Sigma_1 = \Sigma$, and $\pi_1 = \pi_0 = 1/2$, then

- if $\mathcal{Y} = \{0, 1\}$ and $\Sigma_0 = \Sigma_1 = \Sigma$, and $\pi_1 = \pi_0 = 1/2$, then
- define $w = \underbrace{S_W^{-1}}_{d \times d}\underbrace{(\bar{X}_1 - \bar{X}_0)}_{d \times 1}$, $S_W =$ pooled cov. estimate
- estimated Bayes classifier is

$$h^*(x) = \begin{cases} 0 & w^T x > m \\ 1 & w^T x < m \end{cases} \qquad \boxed{m = \frac{1}{2}(\bar{X}_0 + \bar{X}_1)}$$

-

$$w = \arg\max_w \frac{w^T S_B w}{w^T S_W w}$$

$w^T S_B w$ is max'd

- $w^T x \in \mathbb{R}$ maximizes between-group variation, relative to within-group variation

LDA

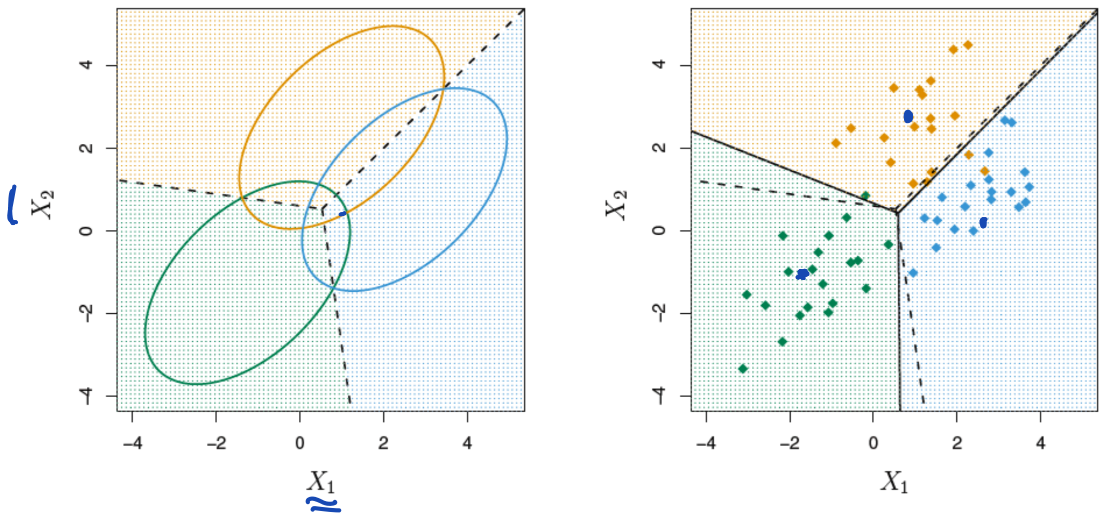


FIGURE 4.6. *An example with three classes. The observations from each class are drawn from a multivariate Gaussian distribution with $p = 2$, with a class-specific mean vector and a common covariance matrix.* Left: *Ellipses that contain 95 % of the probability for each of the three classes are shown. The dashed lines are the Bayes decision boundaries.* Right: 20 *observations were generated from each class, and the corresponding LDA decision boundaries are indicated using solid black lines. The Bayes decision boundaries are once again shown as dashed lines.*

$\Sigma_1 = \Sigma_2 = \Sigma_3$

Error rates

AoS Ex 22.8

- confusion matrix

	classified as 0	classified as 1
$y = 0$	277	25
$y = 1$	116	44

MASS :: lda
qda

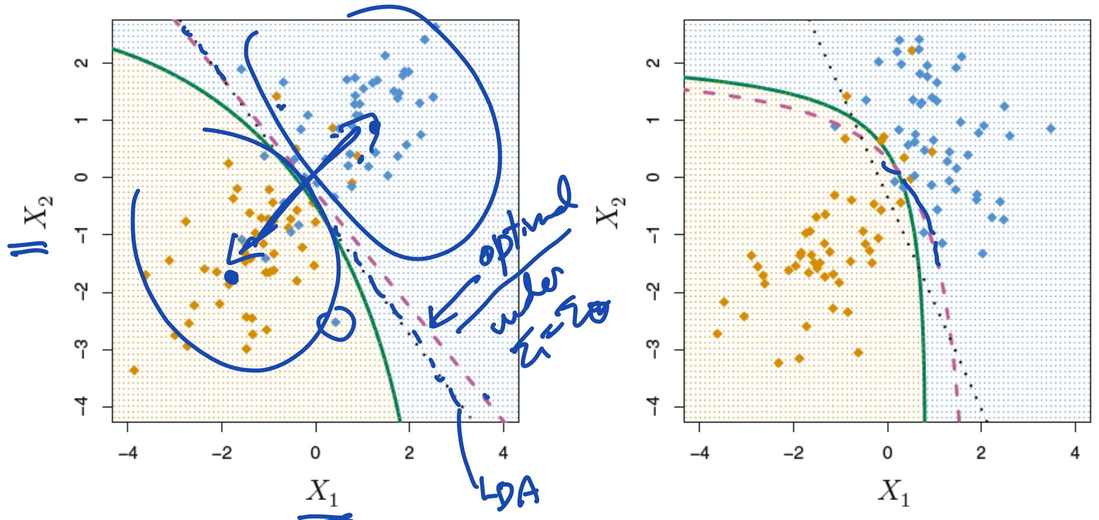
$r(x) \gtrless \frac{1}{2}$

- misclassification rate $(\underline{25 + 116})/\underbrace{(25 + 116 + 277 + 44)} = 0.31$
- see ISLR §4.4 for discussion of changing cut-off from $1/2$ to other thresholds
- changing the loss function to be asymmetric

QDA example

ISLR 4.4

150 4. Classification



$\Sigma_0 \neq \Sigma_1$

FIGURE 4.9. Left: *The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with* $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. *The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA.* Right: *Details are as given in the left-hand panel, except that* $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. *Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.*

Logistic regression

- Model distribution of Y, given X

- $$\mathrm{pr}(Y_i = 1 \mid x_i) = \frac{\exp(\beta_0 + x_i^T\beta)}{1 + \exp(\beta_0 + x_i^T\beta)}$$

- Model distribution of Y, given X
- $$\mathrm{pr}(Y_i = 1 \mid x_i) = \frac{\exp(\hat\beta_0 + x_i^T \hat\beta)}{1 + \exp(\hat\beta_0 + x_i^T \hat\beta)}$$

$\hat\beta_j \quad \hat{se}(\hat\beta_j)$

$y_0 \; 1, \; 0 \qquad \hat p(y_0)\ \ \hat p(x_0) \gtrless \frac{1}{2}$

- equivalently

$$\log\left(\frac{\mathrm{pr}(Y_i = 1 \mid x_i)}{\mathrm{pr}(Y_i = 0 \mid x_i)}\right) = \underline{\beta_0 + x_i^T \beta}$$

$\hat\beta$ est'd

- compare LDA

$$\log\left(\frac{\mathrm{pr}(Y_i = 1 \mid x_i)}{\mathrm{pr}(Y_i = 0 \mid x_i)}\right) = \underline{\alpha_0 + x_i^T \alpha}$$

$$\boxed{\alpha^T = (\mu_1 - \mu_0)^T \Sigma^{-1}}$$

$\hat\alpha$ est'd $\quad f_x(x; \mu, \Sigma)$

Naive Bayes

- K-class again:

$$h(x) = \arg\max_k \mathrm{pr}(Y = k \mid X = x) = \arg\max_k \pi_k f_k(x)$$

$\hat{h}(x) = \hat{\pi}_k \hat{f}_k(x)$

- we could estimate $f_k(x)$ instead of assuming Gaussian, but $X \in \mathbb{R}^d$ curse of dimensionality
- pretend X_j are independent $j = 1, \ldots, d$
- one-dim density estimates from class k:

$$\hat{f}_k(x) = \prod_{j=1}^{d} \hat{f}_{kj}(x)$$

- class probabilities

$$\hat{\pi}_k = \frac{1}{n} \sum 1\{Y_i = k\}$$

- classifier

$$h(x) = \arg\max_k \hat{\pi}_k \hat{f}_k(x)$$

Error rates

- empirical error rate

misclassification rate

$$\hat{L}_n(h) = \frac{1}{n}\sum_{i=1}^{n} 1\{h(X_i) \neq y_i\}$$

training set error

- increasing dimension of X_i decreases training error

e.g. adding variables in logistic regression

- true error rate

$$L(h) = \mathrm{pr}\{h(X) \neq Y\}$$

$f_{X,Y}(x,y)$

- test error rate

$$L(\hat{h}) = \mathrm{pr}\{\hat{h}(X_0) \neq Y_0 \mid \mathcal{T}\}$$

$(X_0, Y_0 \mid \underset{\sim}{X}_n\ \underset{\sim}{Y}_n)$

training

- average test error rate

$$\mathrm{E}_{\mathcal{T}}\{L(\hat{h})\}$$

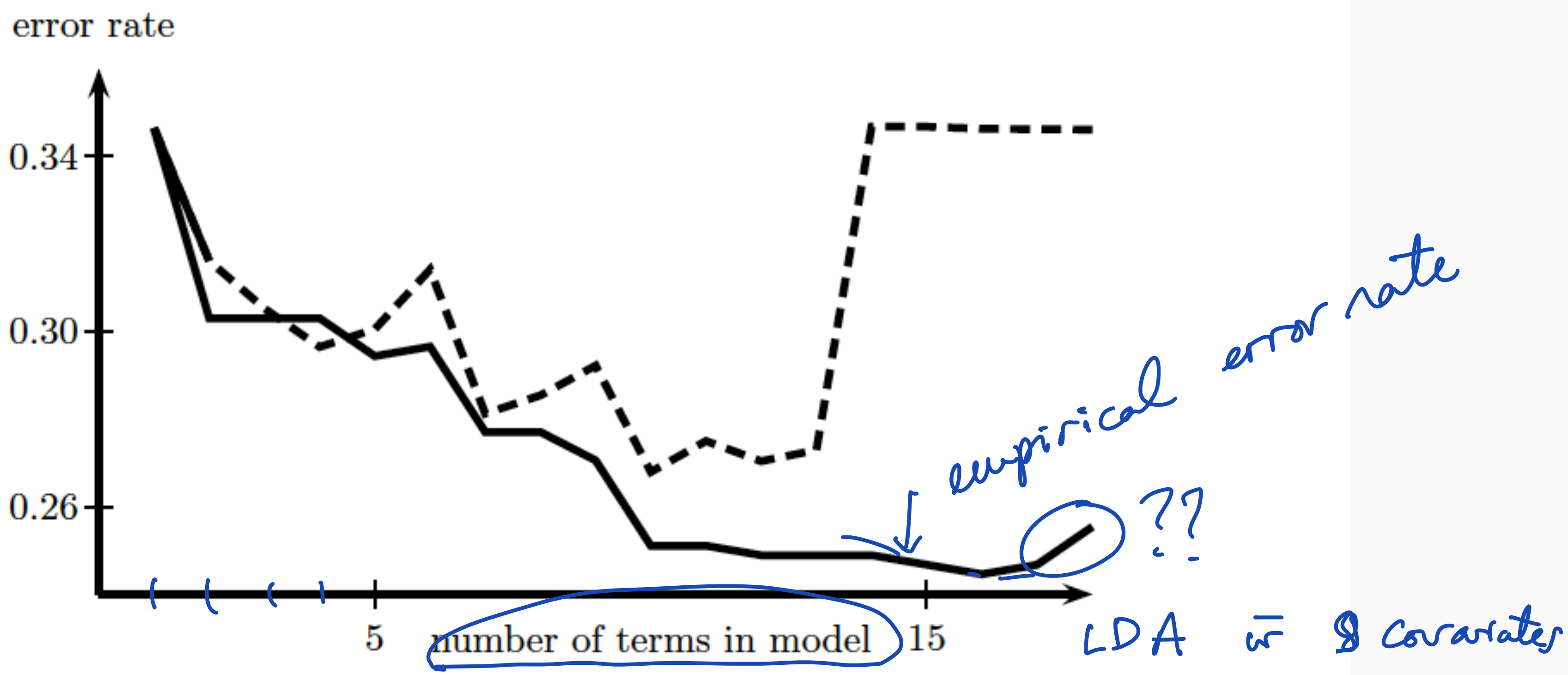


FIGURE 22.5. The solid line is the observed error rate and dashed line is the cross-validation estimate of true error rate.

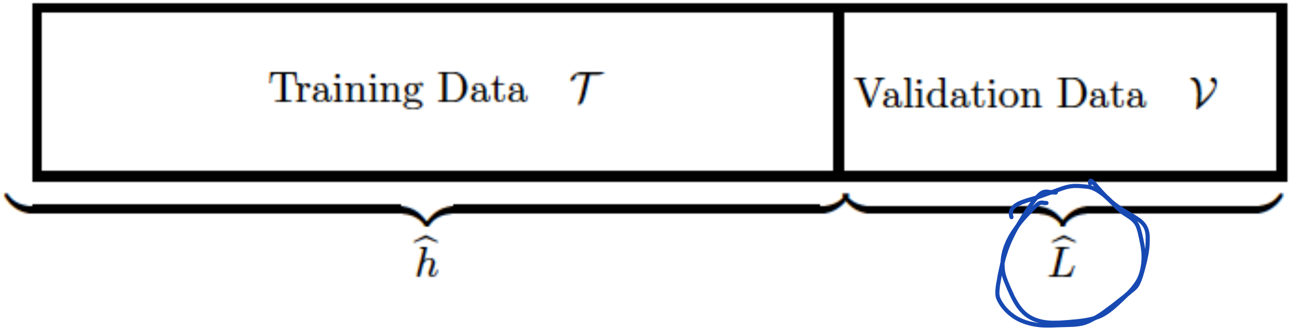


FIGURE 22.6. Cross-validation. The data are divided into two groups: the training data and the validation data. The training data are used to produce an estimated classifier $\widehat{h}$. Then, $\widehat{h}$ is applied to the validation data to obtain an estimate $\widehat{L}$ of the error rate of $\widehat{h}$.

- estimate test error rate with

$$\hat{L}_m(h)\frac{1}{m}\sum_{i\in\mathcal{V}} 1\{\hat{h}(X_i) \neq Y_i\}$$

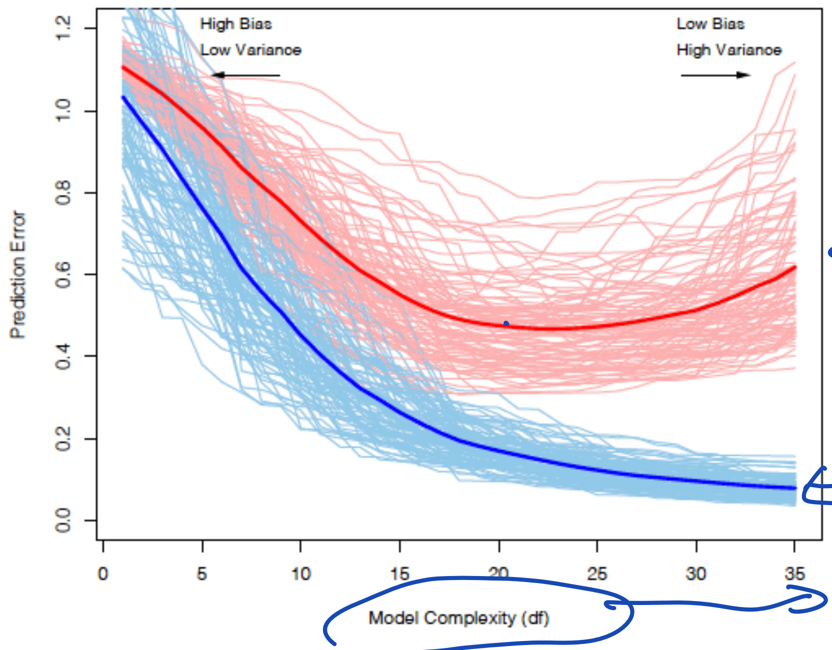
Cross-validation

K-fold cross-validation.

1. Randomly divide the data into K chunks of approximately equal size. A common choice is $K = 10$.
2. For k = 1 to K, do the following:
   (a) Delete chunk k from the data.
   (b) Compute the classifier $\widehat{h}_{(k)}$ from the rest of the data.
   (c) Use $\widehat{h}_{(k)}$ to the predict the data in chunk k. Let $\widehat{L}_{(k)}$ denote the observed error rate.
3. Let

$$\widehat{L}(h) = \frac{1}{K} \sum_{k=1}^{K} \widehat{L}_{(k)}. \tag{22.33}$$

Error rates



test

~~test~~ training

FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\mathrm{err}}$*, while the light red curves show ~~the conditional~~ test error* $\mathrm{Err}_{\mathcal{T}}$ *for 100 training sets of size 50 each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\mathrm{E}[\overline{\mathrm{err}}]$.

Other classification methods

AoS 22.7,9,12; ISLR 2.2

ISLR 2.2

- KNN: K– nearest neighbours
- choose a distance measure on $\mathcal{X}$
- estimate $\mathrm{pr}(Y_1 \mid x)$ by averaging over “nearest neighbours” of x

40 2. Statistical Learning

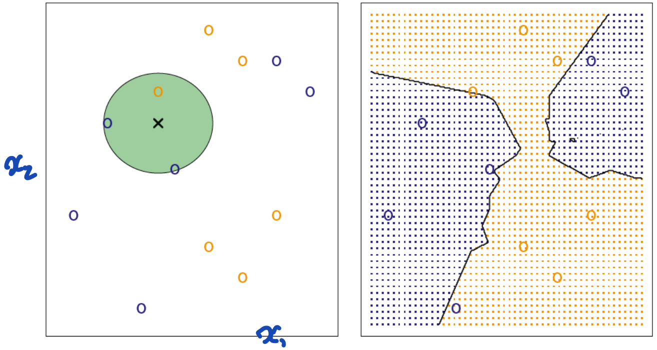
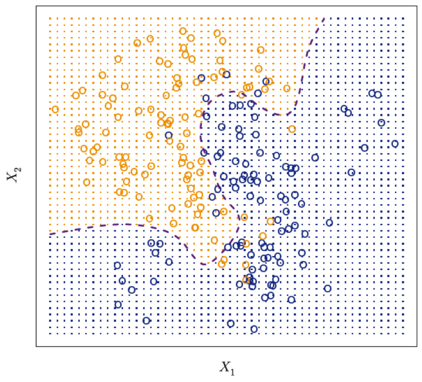


FIGURE 2.14. *The KNN approach, using $K = 3$, is illustrated in a simple situation with six blue observations and six orange observations.* Left: *a test observation at which a predicted class label is desired is shown as a black cross. The three closest points to the test observation are identified, and it is predicted that the test observation belongs to the most commonly-occurring class, in this case blue.* Right: *The KNN decision boundary for this example is shown in black. The blue grid indicates the region in which a test observation will be assigned to the blue class, and the orange grid indicates the region in which it will be assigned to the orange class.*

38 2. Statistical Learning



argmax $f_k(x)\pi_k$

FIGURE 2.13. *A simulated data set consisting of 100 observations in each of two groups, indicated in blue and in orange. The purple dashed line represents the Bayes decision boundary. The orange background grid indicates the region in which a test observation will be assigned to the orange class, and the blue background grid indicates the region in which a test observation will be assigned to the blue class.*

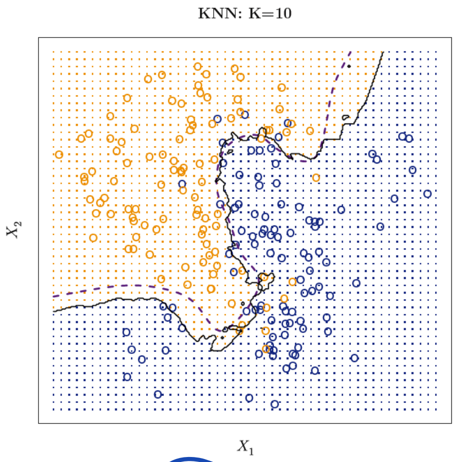


FIGURE 2.15. *The black curve indicates the KNN decision boundary on the data from Figure 2.13, using $K = 10$. The Bayes decision boundary is shown as a purple dashed line. The KNN and Bayes decision boundaries are very similar.*

Classification trees

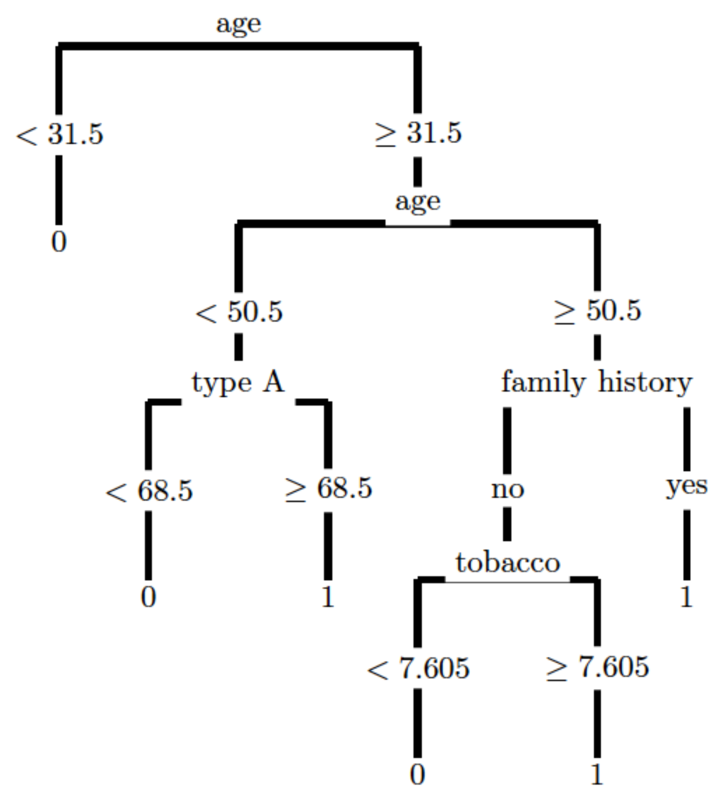


FIGURE 22.7. Smaller classification tree with size chosen by cross-validation.

Other classification methods

- tree-based methods: bagging, boosting, random forests — ISLR 8; ELSII 10
- support vector machines; kernelized SVMs — ISLR 9; ESLII 12
- smoothing logistic regression — ISLR 7.7.2
- neural networks — ELSII 13
- double-descent in deep learning — link