Chapters 1 & 2. Introduction & Overview

Wei Pan

Division of Biostatistics, School of Public Health, University of Minnesota,
Minneapolis, MN 55455
Email: weip@biostat.umn.edu

PubH 7475/8475

Big Data

- *Big Data is on the rise, bringing big questions* (WSJ, 11-29-2012)
  just try a Google search on "Big Data"
- *Big data: the next frontier for innovation, competition, and productivity* (McKinsey report 05-2011)
  from a business perspective, that an enterprise mine all the data it collects right across its operations to unlock golden nuggets of business intelligence (WSJ, 04-29-2012).
- *Big Data's big problem: little talent* (WSJ, 04-29-2012)
  "though bits of it do exist in various university departments and businesses, as an integrated discipline it is only just starting to emerge".
- Recent NSF, NIH Big Data initiatives; NIH PMI.
  2014 NIH Big Data RFA: needs CS, Stat/Math, bio.
- Projects/platforms: CancerLinQ; IBM Watson (Health) ...

- How is this related to statistics?
- Change and expand the subjects
  Many unhappy with the current culture (Breiman, Hand, ...);
  "Data Science" (Cleveland 2001/2014; Yu 2014);
  Computing: Hadoop (or RHadoop), MapReduce, Spark, ...
- You do not need to do everything ...
  DeltaRho (formerly, Tessera): interface b/w R and Hadoop...
  `http://deltarho.org/`
  R packages `datadr`, `trelliscope`
  Based on "Divide and Recombine" (D&R) (Guha et al 2012).
- So ...still need to go back to the basics of ...!

Introduction

- Focus: prediction or discovery.
  Approach: build a model $\hat{f}(x)$.
- Types: supervised vs unsupervised vs semi-supervised learning.
  Training data: with vs without known response values vs a mixture of both.
- Supervised learning: classification vs regression.
  Training data: (Y_i, X_i)'s; Y_i is categorical (e.g. binary) vs quantitative.
  X_i: typically multivariate and mixed types.
  Tuning and test data: (Y_i, X_i)'s;
  Future use: only X_i's.

Examples

- Example 1. X_i^0: an email; $Y_i = 0$ or 1, indicating whether it is a junk email; $i = 1, ..., 4601$.
- Feature extraction: e.g. use some key words in emails as X_i.
- A classification problem: use a 0-1 loss, build a model $\hat{f}(x) \in \{0, 1\}$, calculate misclassification rate,...
- Loss function: here a false positive is much more costly than a false negative.

- Example 2. Predict prostate specific antigen (PSA) using some lab measurements.
- A regression problem.
- Example 3. Handwritten digit recognition.
- X_i^0: a 16 by 16 black/white image ($=$ a 16 by 16 binary matrix); $Y_i \in \{1, 2, ..., 9\}$.
- X_i: maybe (vectorized) X_i^0, or better its summary stat's, e.g. marginal histograms or numbers of "crossing changes" ...

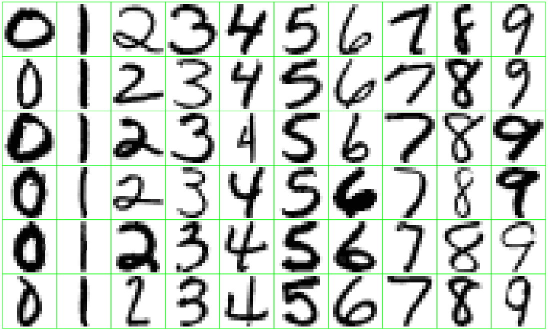
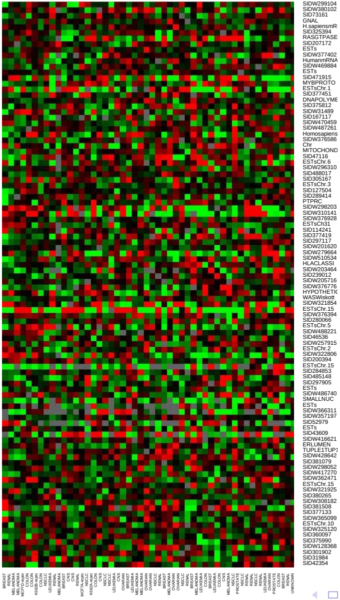


FIGURE 1.2. *Examples of handwritten digits from U.S. postal envelopes.*


SIDW299104
SIDW380102
SID73161
GNAL
H.sapiensmR
SID325394
RASGTPASE
SID207172
ESTs
SIDW377402
HumanmRNA
SIDW469884
ESTs
SID471915
MYBPROTO
ESTsChr.1
SID377451
DNAPOLYME
SID375812
SIDW31489
SID167117
SIDW470459
SIDW487261
Homosapiens
SIDW376586
Chr
MITOCHOND
SID47116
ESTsChr.6
SIDW296310
SID488017
SID305167
ESTsChr.3
SID127504
SID289414
PTPRC
SIDW298203
SIDW310141
SIDW376928
ESTsCh31
SID114241
SID377419
SID297117
SIDW201620
SIDW279664
SIDW510534
HLACLASSI
SIDW203464
SID239012
SIDW205716
SIDW376776
HYPOTHETIC
WASWiskott
SIDW321854
ESTsChr.15
SIDW376394
SID280066
ESTsChr.5
SIDW488221
SID46536
SIDW257915
ESTsChr.2
SIDW322806
SID200394
ESTsChr.15
SID284853
SID485148
SID297905
ESTs
SIDW486740
SMALLNUC
ESTs
SIDW366311
SIDW357197
SID52979
ESTs
SID43609
SIDW416621
ERLUMEN
TUPLE1TUP1
SIDW428642
SID381079
SIDW298052
SIDW417270
SIDW362471
ESTsChr.15
SIDW321925
SID380265
SIDW308182
SID381508
SID377133
SIDW365099
ESTsChr.10
SIDW325120
SID360097
SID375990
SIDW128368
SID301902
SID31984
SID42354
BREAST
RENAL
MELANOMA
MELANOMA
MCF7D-repro
COLON
COLON
K562B-repro
COLON
NSCLC
LEUKEMIA
RENAL
MELANOMA
BREAST
CNS
CNS
RENAL
MCF7A-repro
NSCLC
K562A-repro
COLON
CNS
NSCLC
NSCLC
LEUKEMIA
CNS
OVARIAN
BREAST
LEUKEMIA
MELANOMA
MELANOMA
OVARIAN
OVARIAN
NSCLC
RENAL
BREAST
MELANOMA
OVARIAN
OVARIAN
NSCLC
RENAL
BREAST
MELANOMA
LEUKEMIA
COLON
BREAST
LEUKEMIA
COLON
CNS
MELANOMA
NSCLC
PROSTATE
NSCLC
RENAL
RENAL
NSCLC
RENAL
LEUKEMIA
OVARIAN
PROSTATE
COLON
BREAST
RENAL
UNKNOWN

- Example 4. Microarray gene expression data.
- X_i: 6830 genes' expression levels; quantitative;
  Y_i: tumor types.
- A typical "smalll n, large p" problem: $n = 64$ vs $p = 6830$.
- A classification problem.
- Can be an unsupervised learning problem: finding subtypes of cancer.
  only use X_i's to find new class labels Y_i^*; clustering analysis.
- Can be a semi-supervised learning problem: some known and possibly novel subtypes of cancer.

Overview

- Consider two popular, yet simple and extreme methods: LR vs NN;
  parametric vs non-parametric.
- Q: Is a non-parametric method better than a parametric one? or reverse?
- Consider simulated data: (Y_i, X_i), $Y_i = 0$ or 1 and X_i bivariate; 100 obs's in each class (as training data).
- LR: $E(Y_i|X_i) = Pr(Y_i = 1|X_i) = \beta_0 + X_i'\beta$;
  Use LS to estimate β's $\Longrightarrow \hat{Y}_i = \widehat{Pr}(Y_i = 1|X_i)$;
  $\tilde{Y}_i = I(\hat{Y}_i \geq 0.5)$.
- Decision boundary: $\hat{Y}(x) = \hat{\beta}_0 + x'\hat{\beta} = 0.5$, linear.

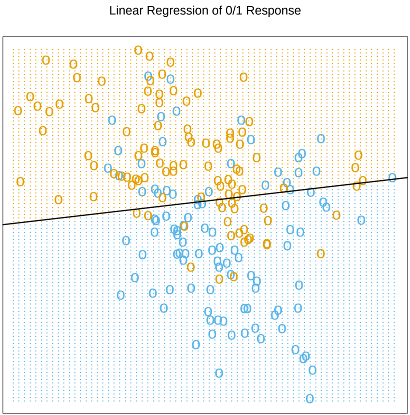


FIGURE 2.1. *A classification example in two dimensions. The classes are coded as a binary variable (BLUE = 0, ORANGE = 1), and then fit by linear regression. The line is the decision boundary defined by* $x^T\hat{\beta} = 0.5$. *The ...*

- kNN: $N_{k(x)}$ is the k nearest training data points that are closest to x,

$$\hat{Y}(x) = \frac{1}{k} \sum_{X_i \in N_k(x)} Y_i = \widehat{Pr}(Y_i = 1 | X_i).$$

- Idea: using local "smoothness" to estimate the population mean by ...
- Key: choice of k, or how much "smoothness" is to be assumed; do not know!
  Modeling assumption: larger k, higher or lower model complexity?
- Try a few values of k, then ...

15-Nearest Neighbor Classifier

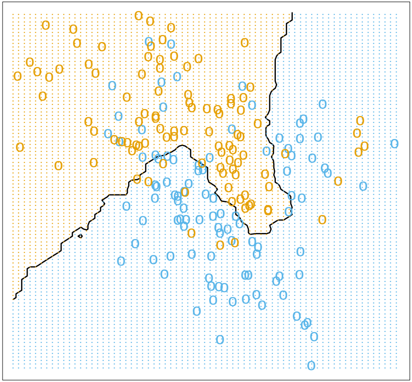


FIGURE 2.2. *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable* (BLUE $= 0$, ORANGE $= 1$) *and then fit by 15-nearest-neighbor averaging as in (2.8). The pre-*

1-Nearest Neighbor Classifier

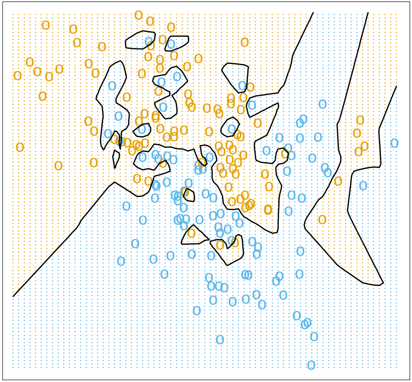


FIGURE 2.3. *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable (*BLUE *= 0,* ORANGE *= 1), and then*

- Key Q: which kNN (and LR) to use?
- Key: cannot use the training data to compare models!
  Why not? too optimistic, favoring ...
  Recall: how to estimate the noise variance in linear regression?
- How? use a separate test dataset, or CV, or some model selection criterion (if any).
  Key: test data should **not** be used in model building!
  Q: how about AIC, BIC ,...
- Previous example: *generate a new test dataset* with $n = 10,000$.

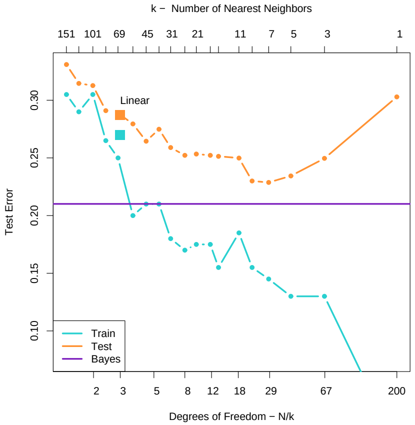


FIGURE 2.4. *Misclassification curves for the simulation example used in Figures 2.1, 2.2 and 2.3. A single training sample of size* 200 *was used, and a test sample of size* 10,000*. The orange curves are test and the blue*

- Q: is there a best classifier?
- Ideal situation: if we know the data distribution, then use the Bayes rule:

$$k_0 = \arg\max_k Pr(k|x).$$

- An example: 1) prior $\pi_k = Pr(k)$; 2) PDF of class k, $f_k(x) = f(x|k)$, then

$$Pr(k|x) = \frac{\pi_k f_k(x)}{\sum_i \pi_i f_i(x)}.$$

  If f_k is assumed to be Normal, then LDA or QDA.
  LR and kNN are also estimating $Pr(k|x)$.
- Bayes rule: offering a theoretical lower bound of the test error rate; often unknown.
- Previous example: R code example 2.1.

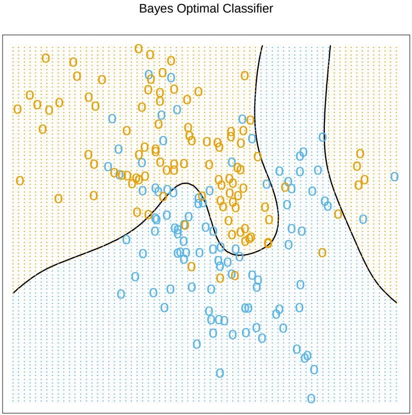


FIGURE 2.5. *The optimal Bayes decision boundary for the simulation example of Figures 2.1, 2.2 and 2.3. Since the generating density is known for each class, this boundary can be calculated exactly (Exercise 2.2).*

- Q: for real data, often cannot generate new data; how to evaluate models?
- Use sample splitting: divide the original whole dataset into two parts, (e.g. 1/2 or 2/3) for training and (the remaining) for test.
  efficient?
- Use cross-validation (CV); read §7.10
- K-fold CV: Divide the data D into almost equally sized and none-overlapping $D_1,...,D_K$, then

$$\text{CVerr} = \sum_{j=1}^{K} \sum_{(Y_i,X_i)\in D_j} L[Y_i, \hat{f}(X_i|D - D_j)]/n.$$

- Leave-One-Out-CV (LOOCV): $K = n$.
- Remarks: 1) not necessarily larger K, the better; CV related to AIC/BIC; 2) maybe better to use bootstrap (§7.11).
- Previous example: R code example 2.1.

- Key: celebrated bias-variance trade-off!
- Suppose $\hat{f}$ is any estimate of f,

$$\begin{aligned} MSE &= E[(\hat{f} - f)^2] = E[(\hat{f} - E(\hat{f}) + E(\hat{f}) - f)^2] \\ &= E[(\hat{f} - E(\hat{f})^2] + E[E(\hat{f}) - f)^2] \\ &= \text{Var} + \text{Bias}^2. \end{aligned}$$

- Very very useful: helps explain
  i) Complex models vs simple models;
  ii) Nonparametrics vs parametrics; ...
- Perhaps the most important plot in the course:

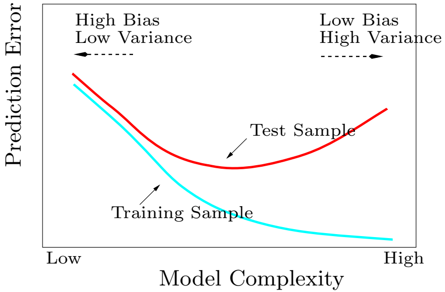


FIGURE 2.11. *Test and training error as a function of model complexity.*