CS 66: Machine Learning

Prof. Sara Mathieson

Spring 2019

Admin

- **Midterm TODAY** (in lab)
  - You may ask questions, but I probably won't say much
  - If you're running low on time: can leave computations un-simplified, shorter written explanations
- **Lab 4 due March 8** (Friday before spring break)

Feedback Form

Feedback form: know well

K-nearest neighbors	18
Decision trees	18
(Linear) regression	14
Probability	2
Likelihoods	1
SGD	2
ML terms, etc	4
Naïve Bayes	1

Feedback form: needs work

• Math behind things	3
• Decision trees	3
• Polynomial regression	5
• Logistic regression	18
• Conditional entropy	2
• Likelihoods	2
• SGD	3
• Naïve Bayes	4

Feedback form: in-class options

	Less	More	As is
Slides	8	10	17
Board work	3	14	18
Group work	4	8	21
Handouts	2	18	15

Feedback form: other

- Review of previous class at beginning of class
- Problem sets
- Slides are too fast, board is good pace
- Math is too fast/confusing, math is okay, many like the math, some want more math
- Some prefer labs with no starter code
- Some want labs with more balance between structured and unstructured
- More context, high-level setup
- Don’t like textbook, supplemental readings better
- More optional theory, extra papers

Outline for February 27

- Naïve Bayes for continuous features
- Evaluation metrics
  - Confusion matrices revisited
  - Precision
  - Recall
  - ROC curves
  - Relationship to probabilistic methods

Outline for February 27

- Naïve Bayes for continuous features
- Evaluation metrics
  - Confusion matrices revisited
  - Precision
  - Recall
  - ROC curves
  - Relationship to probabilistic methods

Recap Naive Bayes

$$p(y=k \mid \vec{x}) \propto \underbrace{p(y=k)}_{\text{prior}} \underbrace{\prod_{j=1}^{p} p(x_j \mid y=k)}_{\text{likelihood under NB assumption}}$$

prediction

$$\hat{y} = \underset{k}{\operatorname{argmax}}\ p(y=k) \prod_{j=1}^{p} p(x_j \mid y=k)$$

Discrete features

$$p(y=k) : \hat{\Theta}_k = \frac{N_k + 1}{n + K}$$

N_k → # train where label is k

n → # train

$$p(x_j = v \mid y=k) : \hat{\Theta}_{j,v,k} = \frac{N_{j,v,k} + 1}{N_k + |f_j|}$$

$N_{j,v,k}$ → # train where feature j = v **and** label is k

$|f_j|$ → # of vals for feature j

$$\boxed{P(A \mid B) = \frac{P(A, B)}{P(B)}}$$

Discriminative vs. Generative

- Regression: discriminative model → finds decision boundary
- Naïve Bayes: generative model → estimates probability distributions

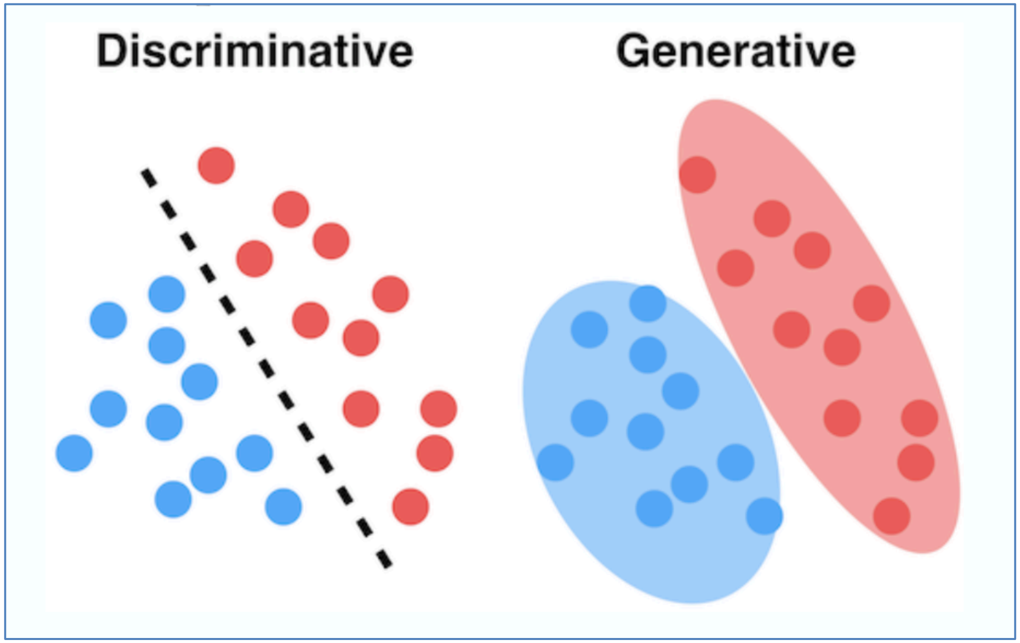


Figure: Ameet Soni

Example with K=2, p=1

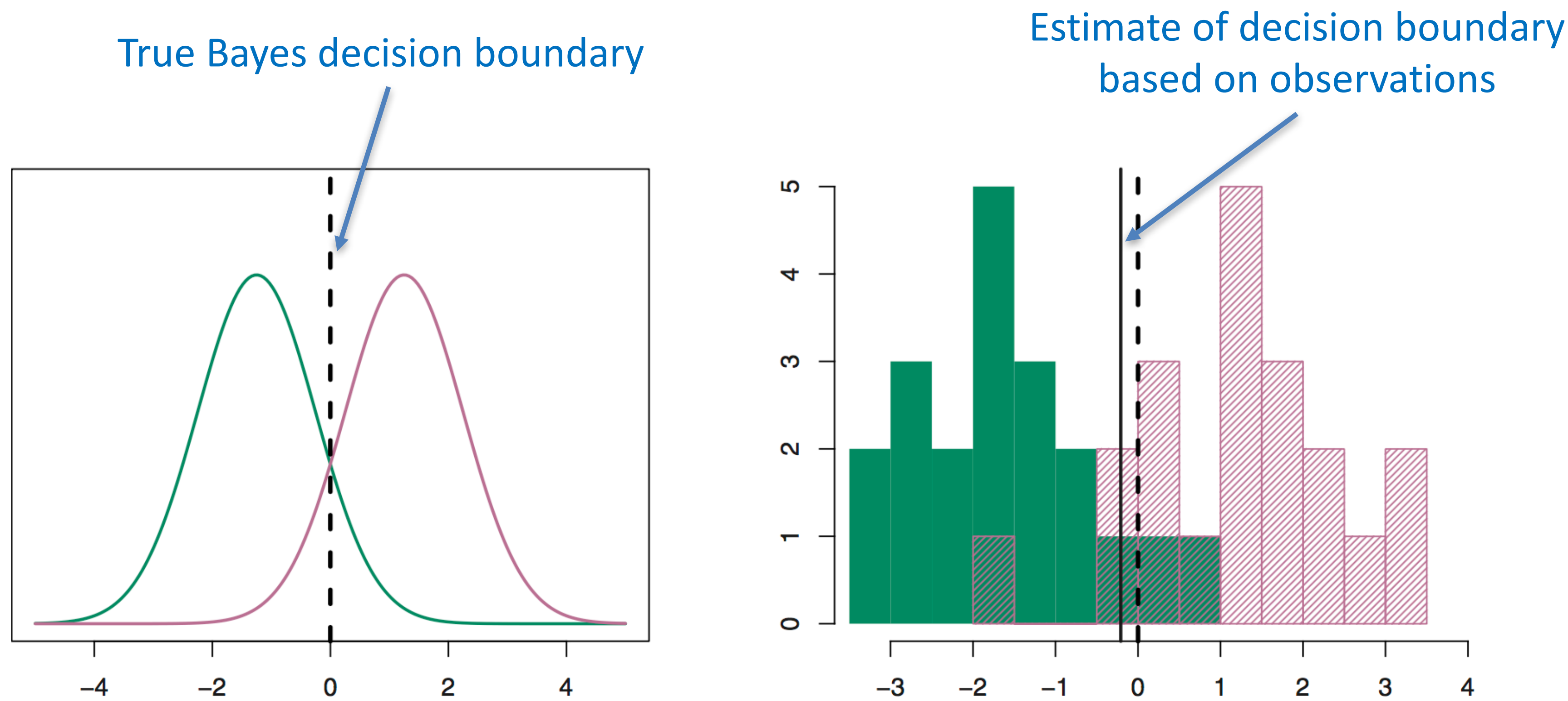


FIGURE 4.4. Left: *Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary.* Right: *20 observations were drawn from each of the two classes, and are shown as histograms.*

Figure: ISL

Example with K=3, p=2

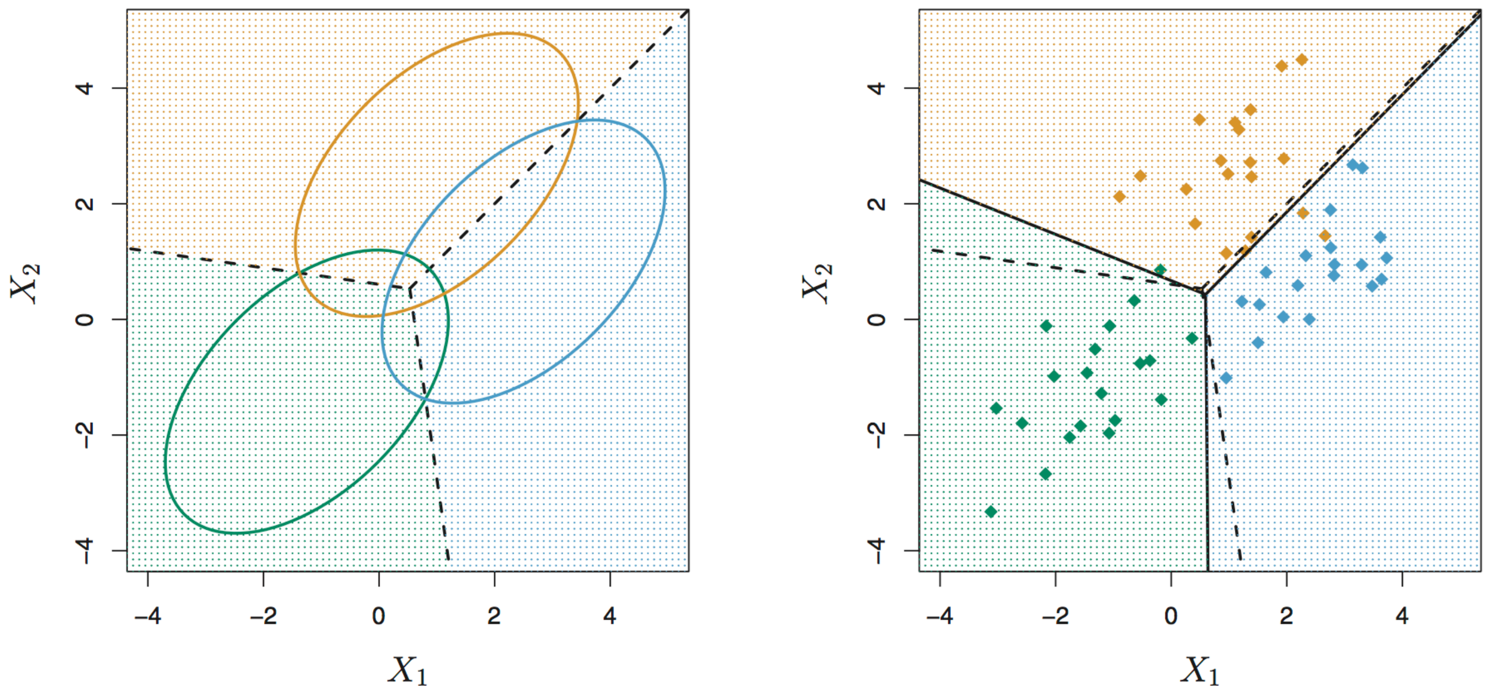
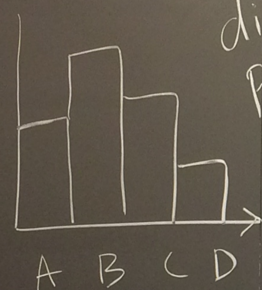


FIGURE 4.6. *An example with three classes. The observations from each class are drawn from a multivariate Gaussian distribution with $p = 2$, with a class-specific mean vector and a common covariance matrix.* Left: *Ellipses that contain* 95 % *of the probability for each of the three classes are shown. 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.*

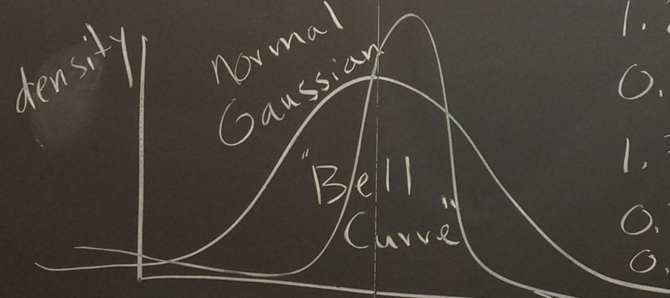
Figure: ISL

Prob



discrete
prob
distribution

$|f_j| = 4$



$p(x_j | y=k)$

$\underbrace{p=1 \qquad K=2}$

X	Y
0.5	1
0.75	1
1.2	2
0.3	1
1.7	2
0.2	2
0.9	2

calcium

thyroid
issue?

p=

p(

$\underline{p=1}$

$$p(x \mid y=k) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)$$

μ_k : mean of $\overbrace{\text{examples}}^{\text{all}}$ with class k

σ^2 : variance of the feature (same for all k)

$$\boxed{\hat{\mu}_k = \frac{1}{N_k} \sum_{i: y_i = k} x_i}$$

$$\hat{\mu}_1 = \frac{0.5 + 0.75 + 0.3}{3}$$

$$\hat{\mu}_2 = \frac{\cdots\cdots}{4}$$

...oid

...ue?

$$\hat{\sigma}^2 = \underbrace{\frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i-\hat{\mu}_k)^2}_{\text{average sample variance}}$$

Outline for February 27

- Naïve Bayes for continuous features
- Evaluation metrics
  - Confusion matrices revisited
  - Precision
  - Recall
  - ROC curves
  - Relationship to probabilistic methods

For now: assume binary classification task

- Transactions that indicate credit card fraud
- Detecting which scans show tumors
- Prenatal test for Down’s Syndrome
- Finding genes under natural selection
- Finding regions of the genome with high recombination rate (“hotspots”)

For now: assume binary classification task

- Transactions that indicate credit card fraud
- Detecting which scans show tumors
- Prenatal test for Down's Syndrome
- Finding genes under natural selection
- Finding regions of the genome with high recombination rate ("hotspots")

In all these examples, we are trying to find unusual items ("needle in a haystack") -- we call these *positives*

true class \ pred	neg	pos	
neg	70	20	90
pos	5	18	

true negative
- TN
- high!

false positive
- FP
- "false alarm"
- Type I error
- want low, but okay sometimes.

false negative
- FN
- "miss"
- want low!
- Type II error

true positive
- want high!
- TP

• precision

$$\frac{TP}{FP+TP}$$ } what fraction of flagged examples are true positives?

$$= \frac{18}{20+18} \approx \boxed{47\%}$$

→ difficult to be high here but that's what we want

• recall

TP

FN

- wa

$b =$

• recall

$$\frac{TP}{FN+TP} = \frac{18}{5+18} \approx \boxed{.78}$$

- want high

$$\vec{b} = (X^T X)^{-1} X^T y$$