Stat 435 Intro to Statistical Machine Learning

Week 3: Homework 1, Probability, Bayes rule, etc.

Richard Li

April 12, 2017

Plan for today

- Recap of Homework 1
- Bayes error rate
  - Definition
  - Simple examples
  - Bayes rule
  - Difficult examples

All but Bayes error rate

Logistics:

- In the future, **don't submit only** .Rmd file.
- Don't forget to "comment" on your findings.

Some other problems:

- 6(c): marginal association v.s. conditional association
- 6(g): In R, *which.min()* and *which.max()* returns only the first result when tie exists. I did not deduct points this time.

All but Bayes error rate

- Deterministic support (problem 2)
    - Does 1-NN overfit?
    - What if we have more training data?

Bayes error rate: definition

Textbook definition:

- Test error rate produced by the **Bayes classifier**.
- **Bayes classifier**: always predict Y to be the class with largest $Pr(Y|x)$.

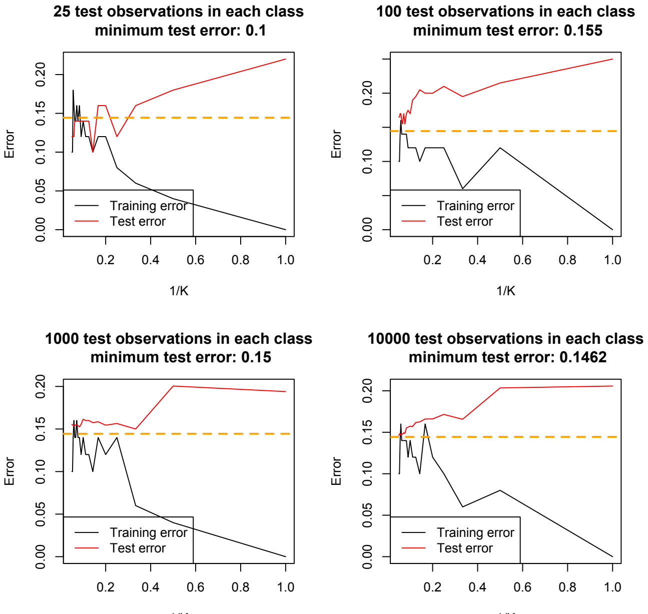
For 2-class problem,

- For any given x, if we can calculate $Pr(Y = 0|x)$ and $Pr(Y = 1|x)$, we always predict Y to be the more likely class.
- What's the risk of doing this?
  - Y could be from the less likely class!
  - For any given x, we expect the error to be $1 - Pr(\text{Y form the most likely class}|x)$.

Bayes error rate: definition

- Bayes error rate: $1 - E(Pr(\text{Y form the most likely class}|x))$.
- The expectation is taken w.r.t. the probability of all possible X.
- It only depends on how X and Y are generated.
- It does **not** depend on training/testing data.
- It is the theoretical lower bound of the *expected error* of any classifier.
- Does **not** guarantee to be lower than any error rates of any classifier **on any test dataset**.

Bayes error rate: HM problem 1 with more test data



Let's look at a few Bayes error rate calculations.

Bayes error rate: deterministic case

In problem 2 of the Homework,

- What is $Pr(\text{Y form the most likely class}|x)$?
- If given any x there is **only one possible** Y, we won't make mistakes.
- So Bayes error rate is 0 in that problem.

In the more general case,

- $0 < Pr(\text{Y form the most likely class}|x) < 1$
- How to calculate Bayes error rate?

Review: Probability density functions

- Discrete case

$$f(x) = Pr(X = x)$$

- Continuous case

$$f(x) = \lim_{\delta \to 0} \frac{Pr(x \leq X \leq x + \delta)}{\delta}$$

Review: Expectation and variance

- Discrete case

$$E(x) = \sum_x xf(x) \quad Var(x) = \sum_x (x - E(X))^2 f(x)$$

- Continuous case

$$E(x) = \int_x xf(x)dx \quad Var(x) = \int_x (x - E(x))^2 f(x)dx$$

Bayes error rate: mathematical form

- Now, Recall Bayes error rate is the error from **Bayes classifier**
- Thus the formula,

$$\text{Bayes error} = 1 - E(\max_j Pr(Y = j|x))$$

- We can also write this as

$$\text{Bayes error} = 1 - \int \max_j Pr(Y = j|x) f(x)dx$$

- Again for two-class problem,

$$\text{Bayes error} = 1 - \Big(\int_{L_0} Pr(Y = 0|x) f(x)dx + \int_{L_1} Pr(Y = 1|x) f(x)dx \Big)$$

- $L_0 = \{x : Pr(Y = 0|x) > 0.5\}$ and $L_1 = \{x : Pr(Y = 1|x) > 0.5\}$

Bayes error rate: simple example

- Suppose the real data are generated such that

$$X \sim Unif[-1, 1]$$

- the true labels (0 or 1) of the data are generated such that

$$Pr(Y = 1|X < 0) = 0.2 \qquad Pr(Y = 1|X > 0) = 0.9$$

- How often do we expect to observe $Y = 1$?

$$\begin{aligned} Pr(Y = 1) &= \int_{-1}^{0} Pr(Y = 1|X < 0) f(x)dx + \int_{0}^{1} Pr(Y = 1|X > 0) f(x)dx \\ &= Pr(Y = 1|X < 0) \int_{-1}^{0} f(x)dx + Pr(Y = 1|X > 0) \int_{0}^{1} f(x)dx \\ &= 0.2 \times 0.5 + 0.9 \times 0.5 = 0.55 \end{aligned}$$

- What is the Bayes error rate?

$$\begin{aligned} Pr(Y = 1) &= \int_{-1}^{0} Pr(Y = 1|X < 0) f(x)dx + \int_{0}^{1} Pr(Y = 0|X > 0) f(x)dx \\ &= 0.2 \times 0.5 + (1 - 0.9) \times 0.5 = 0.15 \end{aligned}$$

Exercise

- What if we change the distribution of X to $X \sim Unif[-10, 1]$?
- What if we change the true labels to be

$$Pr(Y = 1|X < 0) = 0.9$$

$$Pr(Y = 1|0 < X < 0.5) = 0.2$$

$$Pr(Y = 1|0.5 < X) = 0.8$$

- What if we change the true labels to be

$$Pr(Y = \{0, 1, \text{re-accommodate}\}|X > 0) = \{0.5, 0.5, 0\}$$

$$Pr(Y = \{0, 1, \text{re-accommodate}\}|X < 0) = \{0.3, 0.5, 0.2\}$$

Review: Bayes rule

This calculation seems straight forward 😎,
but why is the homework problem so difficult?

Because $Pr(Y|X)$ is not given directly 😕

Bayes rule

- A simple application of *conditional probability*.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

- In fact, this is something you will see extensively in Chap 4 (LDA).

Bayes rule: example

- You have a (faulty) alarm at home, it goes off
  - with probability 0.9 if your home is burglarized;
  - with probability 0.05 if your home is not burglarized...
- Now, your friend calls you and says your alarm just went off!
- What is the probability of your home being burglarized?

Bayes rule: example

- We can write $Pr(A|B=1) = 0.9$, and $Pr(A|B=0) = 0.05$.
- We want to know $Pr(B=1|A)$.
- Use Bayes rule,

$$Pr(B|A) = \frac{Pr(A|B)Pr(B)}{Pr(A)}$$

- Suppose you live in a neighborhood where $P(B=1) = 0.2$,
- Then

$$Pr(B=1|A) = \frac{0.9 \times 0.2}{Pr(A)} \quad Pr(B=0|A) = \frac{0.05 \times 0.8}{Pr(A)}$$

Bayes rule: example

- But we do not know $Pr(A)$ (at least not directly)!
- However, since we know $Pr(B=0|A) + Pr(B=0|A) = 1$
- *We only need the relative proportion of the two*

$$Pr(B=1|A) \propto 0.9 \times 0.2 = 0.18$$

$$Pr(B=0|A) \propto 0.05 \times 0.8 = 0.04$$

- To calculate exact numbers,

$$Pr(B=1|A) = \frac{0.18}{0.18+0.04} \approx 0.82$$

$$Pr(B=0|A) = \frac{0.04}{0.18+0.04} \approx 0.18$$

Bayes error: Homework revisited

- What is the Bayes classifier here?
- How to approximate the Bayes error (through simulation)?
- How to actually calculate it?
- Why my testing error beats the "best" classifier?

Bayes error: Homework revisited

First, what is the Bayes classifier here?

- When you have an x_0, Bayes classifier will assign
  - the class with higher pdf, $f(x_0|y)$;
  - the class whose center ($[0,0]$ or $[1.5,1.5]$) is closer to x_0;
  - the class with higher $Pr(y|x_0)$
- Why are they equivalent?
- What assumptions we are making?

Bayes error: Homework revisited

From the homework solutions on Canvas:

Notice in the algorithm above, we are comparing $f(\mathbf{x}|y)$ instead of $f(y|\mathbf{x})$. There is a good reason why they are equivalent. We know that using Bayes rule, we have

$$Pr(Y=j|x) = \frac{Pr(x|Y=j)Pr(Y=j)}{Pr(x)}$$

We do not know $Pr(x)$, but we can plug in $Pr(Y=j) = 1/2$, and that $Pr(Y=1|x) + Pr(Y=2|x) = 1$. Combining these we can derive the formula that we can actually calculate:

$$Pr(Y=j|x) = \frac{Pr(x|Y=j)}{Pr(x|Y=red) + Pr(x|Y=blue)}$$

Bayes error: Homework revisited

How to get $1 - E(max_j Pr(Y = j|x))$ without doing calculus?

```r
Nsim <- 1e5
total_max_prob <- 0
for(i in 1:Nsim){
    y <- sample(c(1, 2), 1)
    if(y == 1){
        x <- rnorm(2, 0, 1)
    }else{
        x <- rnorm(2, 1.5, 1)
    }
    p1 <- standard_binormal_density(x, c(0, 0))
    p2 <- standard_binormal_density(x, c(1.5, 1.5))
    # use formula
    total_max_prob <- total_max_prob + max(p1 /(p1 + p2), p2 / (p1 + p2))
}
1 - total_max_prob / Nsim
```

```
## [1] 0.1444253
```

Bayes error: Homework revisited

Double check the formula is correct using definition?

```
# use definition
if((p1 < p2 && y == 1) || (p1 > p2 && y == 2)) {
    total_error <- total_error + 1
}
```

Bayes error: Analytical solution (FYI)

To get an analytical solution, we do not resolve to the computational trick of $p_1/(p_1 + p_2)$ as before. Instead by using the original form,

$$E_x(max_j Pr(Y = j|\mathbf{x})) = \int_{R^2} max\{\frac{0.5f_0(\mathbf{x})}{Pr(\mathbf{x})}, \frac{0.5f_1(\mathbf{x})}{Pr(\mathbf{x})}\} Pr(\mathbf{x}) d\mathbf{x}$$

Notice the term $Pr(\mathbf{x})$ cancels out!

$$E_x(max_j Pr(Y = j|\mathbf{x})) = \int_{R^2} max\{0.5f_0(\mathbf{x}), 0.5f_1(\mathbf{x})\} d\mathbf{x}$$

We can calculate the regions where $f_0 < f_1$ and vice versa by observing

$$f_0(x) = f_1(x) \implies x_1^2 + x_2^2 = (x_1 - 1.5)^2 + (x_2 - 1.5)^2$$

Bayes error: Analytical solution (FYI)

$$error = 1 - \left(0.5 \int_{x_2 < 1.5 - x_1} f_0(\mathbf{x}) d\mathbf{x} + 0.5 \int_{x_2 > 1.5 - x_1} f_1(\mathbf{x}) d\mathbf{x}\right)$$

- The integral can be calculated numerically,

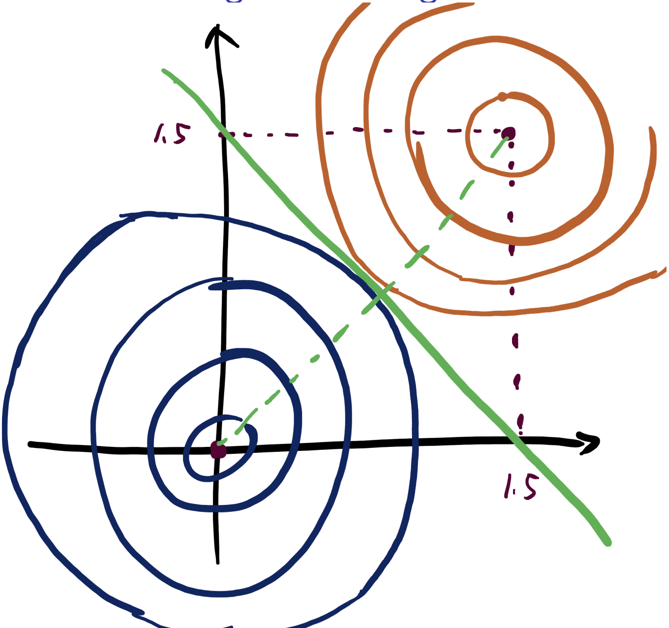
Bayes error: Analytical solution (FYI)

- The integral can be calculated numerically,
- or we can simplify it a little further...

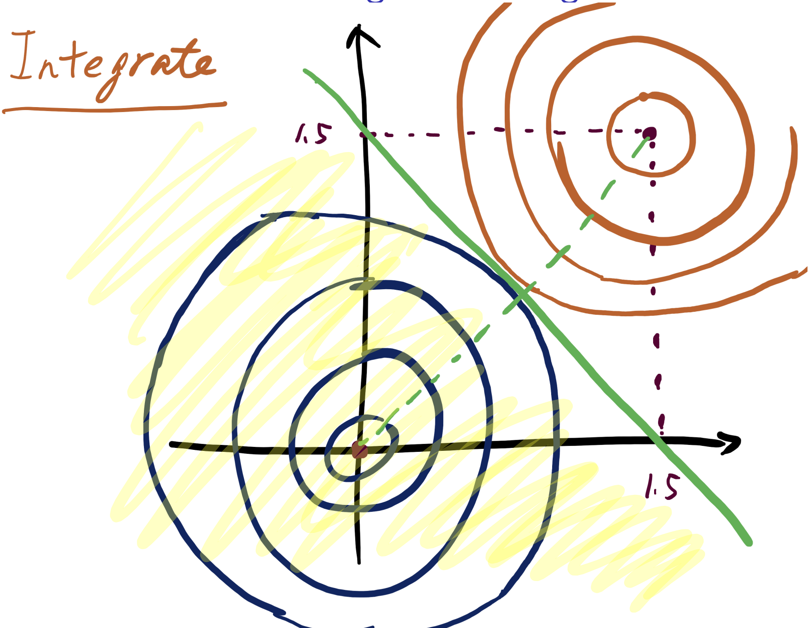
$$\begin{aligned}\int_{x_2<1.5-x_1} f_0(\mathbf{x})d\mathbf{x} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\frac{\sqrt{1.5^2+1.5^2}}{2}} f_0(\mathbf{x})dx_1dx_2 \\ &= \Phi(\frac{\sqrt{1.5^2+1.5^2}}{2})\end{aligned}$$

- Same for the other term.
- $1 - 0.5 * (pnorm(\frac{\sqrt{1.5^2+1.5^2}}{2}) + pnorm(\frac{\sqrt{1.5^2+1.5^2}}{2}))$

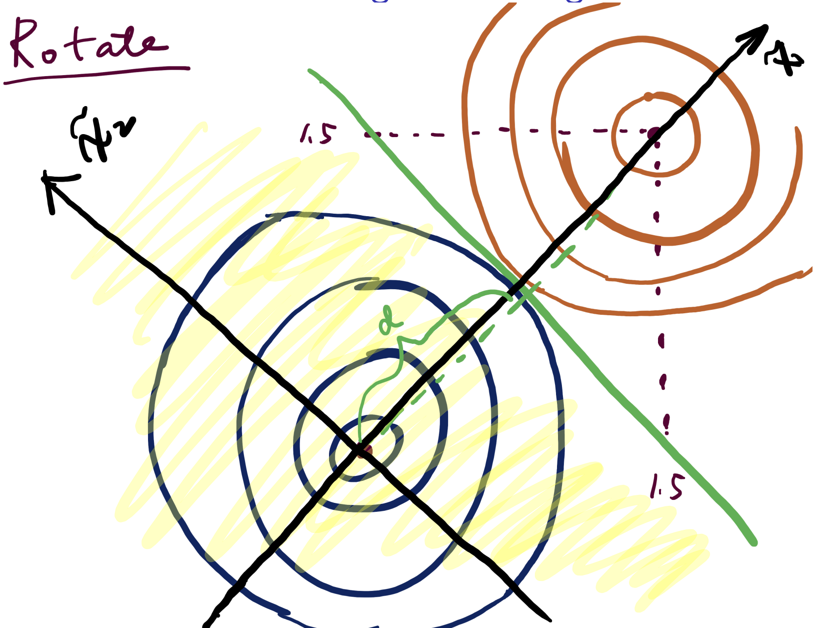
A not so great drawing



A not so great drawing



A not so great drawing



Summary

- Bayes error calculation when
  1. $Pr(Y|X)$ is known
  2. $Pr(X|Y)$ is known
- In the later case, numerical approximation/simulation using
  1. definition of Bayes classifier
  2. formula of Bayes error rate

Why we learn this

- Bayes error rate is great.
- But it can only be calculated if you know the ground truth.
- In practice, we do not have know $P(X|Y)$, $P(X)$, or even $P(Y)$.
- One of the approaches in real-life classification:
  - Assume some generating distribution.
  - Estimate the distribution from data.
  - Making classifications using the 'approximated' Bayes classifier.
- In the Normal $P(X|Y)$ case, this is called LDA.

Why we learn this

4.4 Linear Discriminant Analysis 143

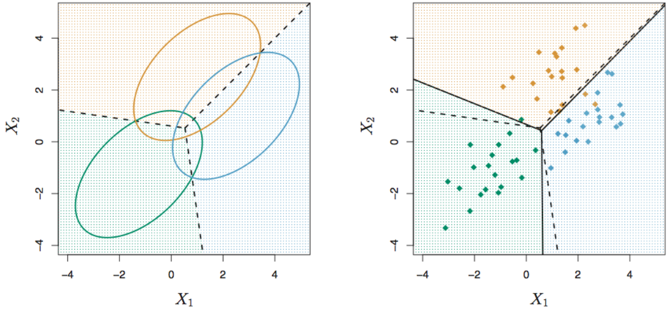


FIGURE 4.6. *An example with three classes. The observations from each class are drawn from a multivariate Gaussian distribution with $p = 2$, with a class-specific mean vector and a common covariance matrix.* Left: *Ellipses that contain 95 % of the probability for each of the three classes are shown.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

Takeaway

"The dragon has three heads: probability rules, calculus, and simulation." – Rhaegar Targaryen, A Clash of Kings.