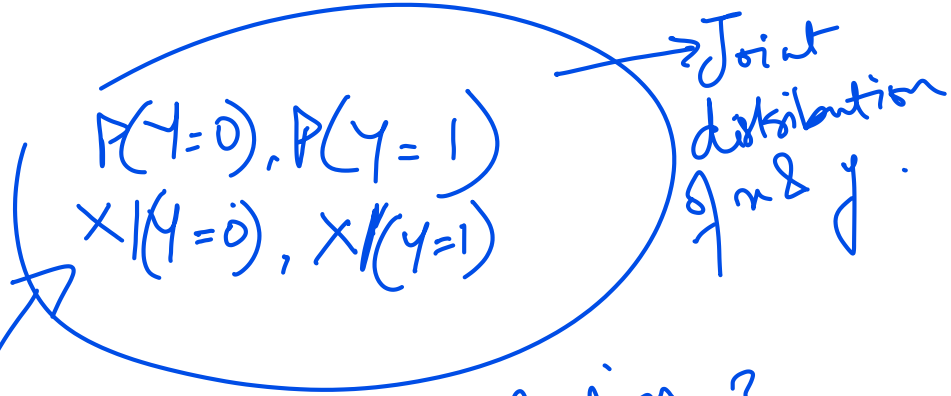
DSO530 Statistical Learning Methods

Lecture 3b: Classification II

Dr. Xin Tong
Department of Data Sciences and Operations
Marshall School of Business
University of Southern California

What is the best classifier?



- It depends on what is your objective function → evaluation metric?
- Suppose your goal is to find h that minimizes $P(h(X) \neq Y)$
- What is the best classifier if you knew the joint distribution of (X, Y)?
- intuitively, how should we think about "know the joint distributin of (X, Y)"?
- Let's take for granted that the best classifier is $1(\eta(x) > 1/2)$, where $\eta(\cdot)$ is the so called regression function.
- This classifier is the so-called `Bayes classifier`
- Recall the regression function: $\eta(x) = E(Y|X = x)$.
- In the binary classification scenario, $E(Y|X = x) = P(Y = 1|X = x)$.

What is the best classifier?

- It depends on what is your objective function
- Suppose your goal is to find h that minimizes $P(h(X) \neq Y)$
- What is the best classifier if you knew the joint distribution of (X, Y)?
- intuitively, how should we think about "know the joint distributin of (X, Y)"? ⟶ sample size ∞
- Let's take for granted that the best classifier is $1(\eta(x) > 1/2)$, where $\eta(\cdot)$ is the so called regression function.
- This classifier is the so-called `Bayes classifier`
- Recall the regression function: $\eta(x) = E(Y|X = x)$.
- In the binary classification scenario, $E(Y|X = x) = P(Y = 1|X = x)$.

Take value 1 if value in parenthesis is True

$$h^*(x) = \mathbb{1}\left(P(Y=1 \mid X=x) > 1/2\right)$$

↳ Indicates optimal classification of minimum error.

Are we done now? No

Since some statisticians have found this best classifier, why don't we just use it and save all the trouble to learn classification methods?

In ML & stats, we start from sample. $\Rightarrow$ we don't know joint distribution of x & y.

We cannot use the Bayes classifier

- Knowing the distribution of (X, Y) is impossible.
- We only know some instances sampled from the distribution.
- So we have to estimate $\eta(x)$ based on the sample
- Where does the logistic regression model come into the picture?

- The logistic regression model is a parametric model for $\eta(x)$ or $P(Y = 1|X = x)$
- But why do we often want to impose such a restrictive form for η?

We cannot use the Bayes classifier

Regression: $Y = f(x) + \varepsilon$, f can be highly complex.

- Knowing the distribution of (X, Y) is impossible.
- We only know some instances sampled from the distribution.
- So we have to estimate $\eta(x)$ based on the sample
- Where does the logistic regression model come into the picture?

There is a limitation to simple models.

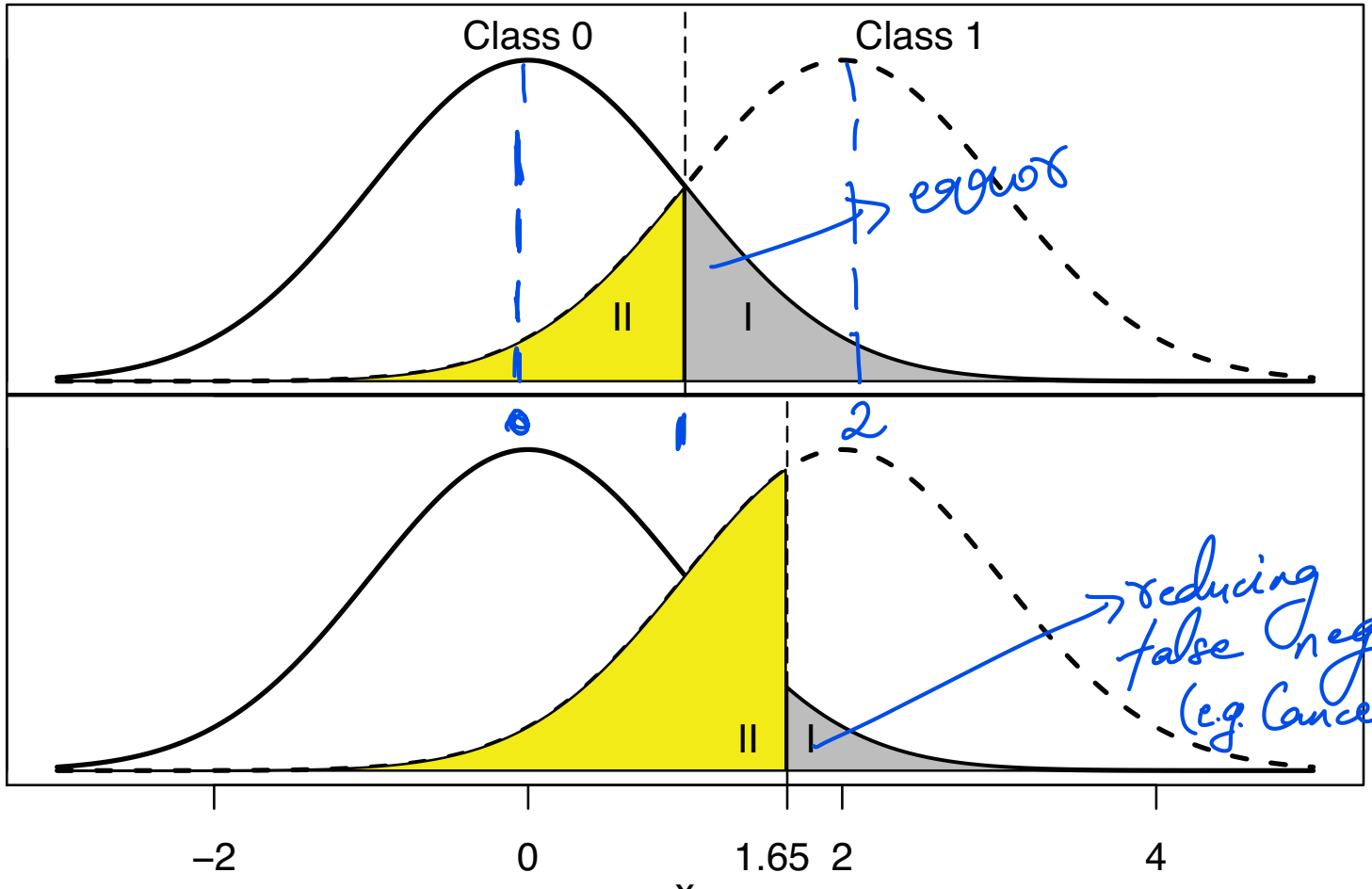
- The logistic regression model is a parametric model for $\eta(x)$ or $P(Y = 1|X = x)$ $P(Y=1|X=x) = e^{\beta_0+\beta_1 x}/(1+e^{\beta_0+\beta_1 x})$
- But why do we often want to impose such a restrictive form for η?

A typical question

$R(h^*) = P(h^*(x) \neq Y) =$ Area I $\cdot 1/2$ + Area II $\cdot 1/2$ → classification error.

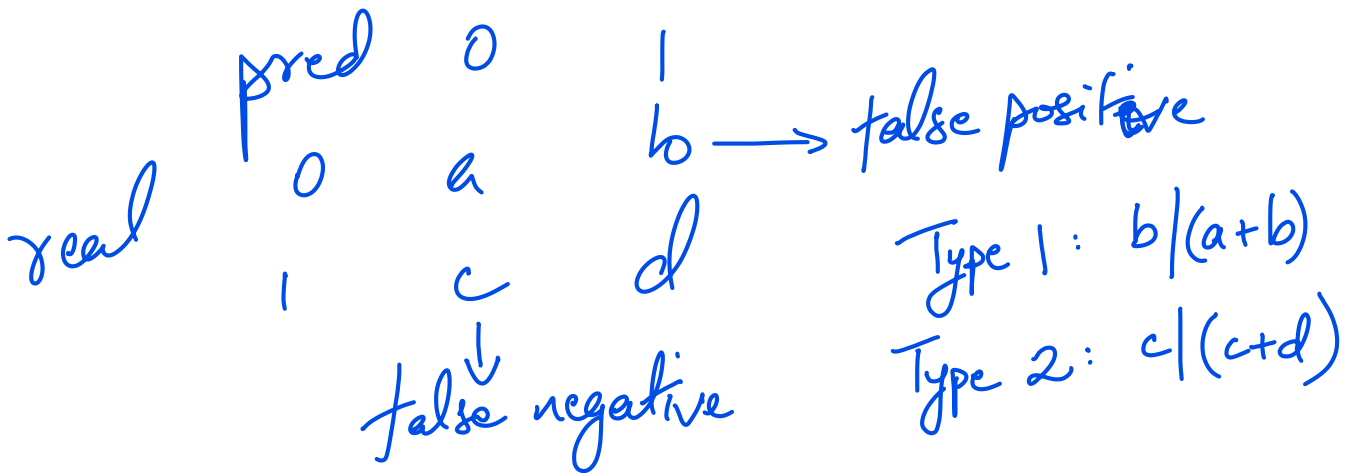
- Q: Getting infinite observations, can I achieve perfect classification?
- A: Typically, no!
- Why: look at the Bayes classifier. Focus on 1st row: class 0: N(0, 1); class 1: N(2, 1); balanced classes.

$x|(Y=1) \sim N(2,1)$ → $h^*(x) = \mathbb{1}(x>1)$

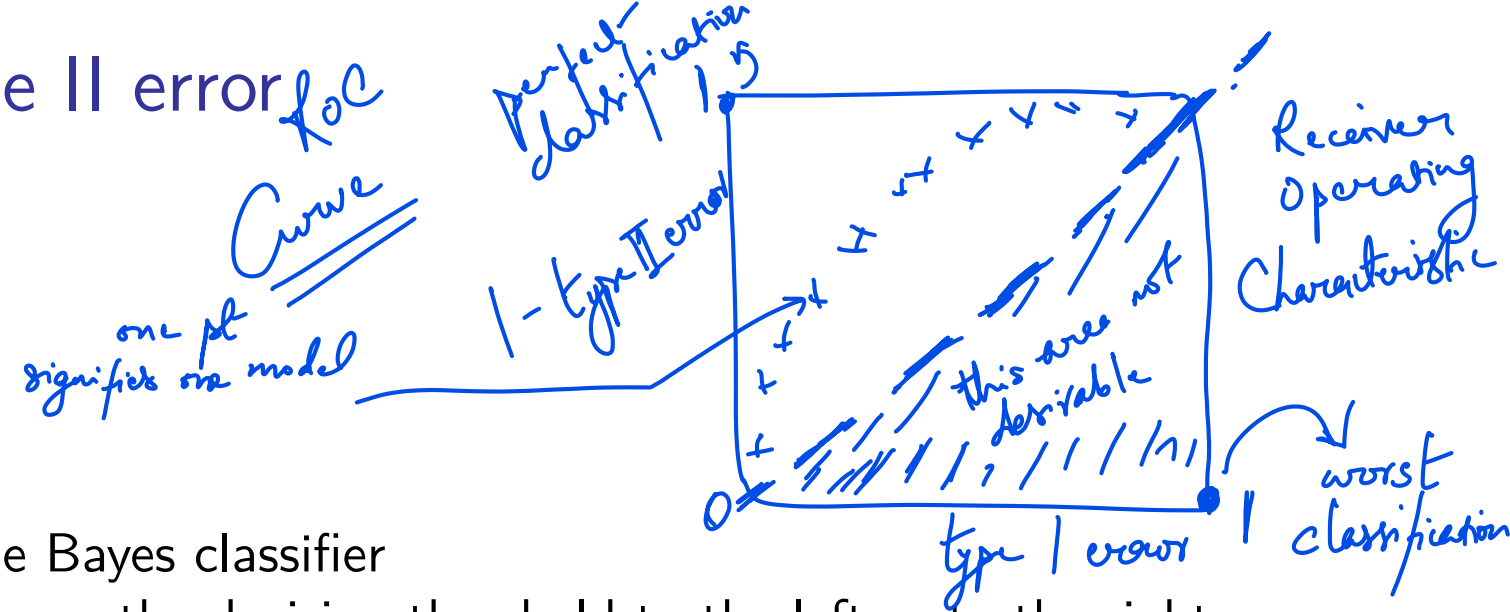


Type I vs. type II error

- Modify the Bayes classifier
- You can move the decision threshold to the left or to the right
- How will type I and type II error change?
- type I error definition: $P(h(X) \neq Y | Y = 0)$ false negative
- type II error definition: $P(h(X) \neq Y | Y = 1)$ false positive
- a takeaway message: we can change the decision threshold so that we rebalance the trade-off between type I and type II errors

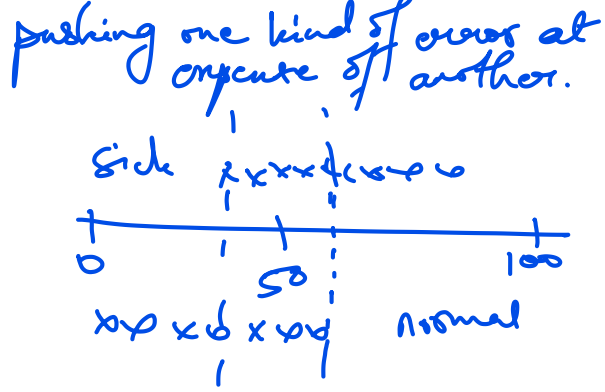


Type I vs. type II error



- Modify the Bayes classifier
- You can move the decision threshold to the left or to the right
- How will type I and type II error change?
- type I error definition: $P(h(X) \neq Y | Y = 0)$
- type II error definition: $P(h(X) \neq Y | Y = 1)$
- a takeaway message: we can change the decision threshold so that we rebalance the trade-off between type I and type II errors

Connection to reality



- Suppose 0 codes disease status and 1 codes normal
- Then type I error is the false negative rate and type II error is the false positive rate
- In the above, we showed that even if you can have the entire instances in the world, you still likely cannot achieve 0% false negative rate and false positive rate.
- Given the current training data and machine learning model, one can push down one kind of error at the expense of the other.
- How can I lower both type I error and type II error at the same time in practice? (1) a better model. (2) enlarge the sample size (3) get more powerful features that can separate the two classes better.
- We should note that the first two solutions have a limit.

terrorist classification example

usually, $X|(Y=0), X|(Y=1)$ have overlaps

- get more data
- find better features which differentiate both class effectively.

A question for you to ponder

Specific example: flip a fair coin (H) → 1, (T) → 0

$P(H) = P(T) = 0.5 \Rightarrow P(\hat{Y}=1) = P(\hat{Y}=0) = 0.5$

type 1 error: misclassifying tails as head. $= P(\hat{Y}=1 \mid Y=0) = 0.5$

type 2 error: misclassifying heads as tails $= P(\hat{Y}=0 \mid Y=1) = 0.5$

0 disease
1 normal

→ Same type 1 error as above but small a type 2 error.

If one gives you a classifier that has false negative rate of 50% and false positive rate of 60%, is it acceptable? No

These cases are possible but not desirable.

Type 1 error: misclassifying class 0 instance

Type 2 error: misclassifying class 1 instance.

↳ because $50\% + 60\% > 1$

flipping biased coin:

$P(\hat{Y}=1) = \alpha$, $P(\hat{Y}=0) = 1-\alpha$

⇒ Type 1 error $= \alpha$

Type 2 error $= 1-\alpha$