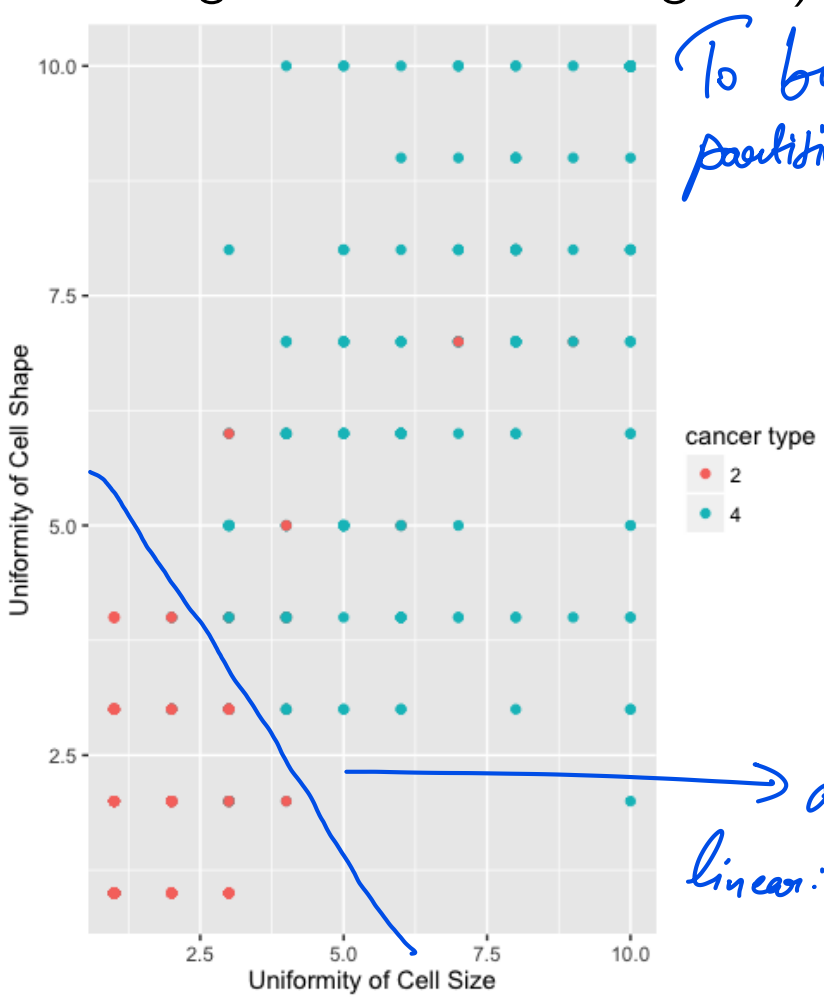
DSO530 Statistical Learning Methods

Lecture 3a: Classification I

Dr. Xin Tong
Department of Data Sciences and Operations
Marshall School of Business
University of Southern California
xint@marshall.usc.edu

Recall a classification example

- This breast cancer dataset was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. (2 codes benign and 4 codes malignant). A classification example



To build a classifier is to have a partition of the feature space.

→ decision boundary.

linear: Logistic regression.

Classification

more popular w/ specific methods, such as Support Vector Machine (SVM)

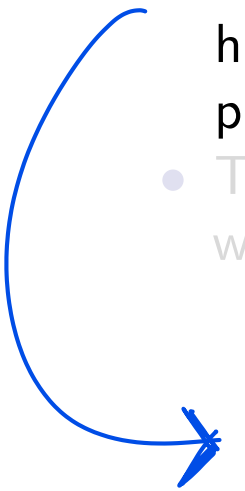
$\{0,1\}$ or $\{-1,1\}$: binary classification

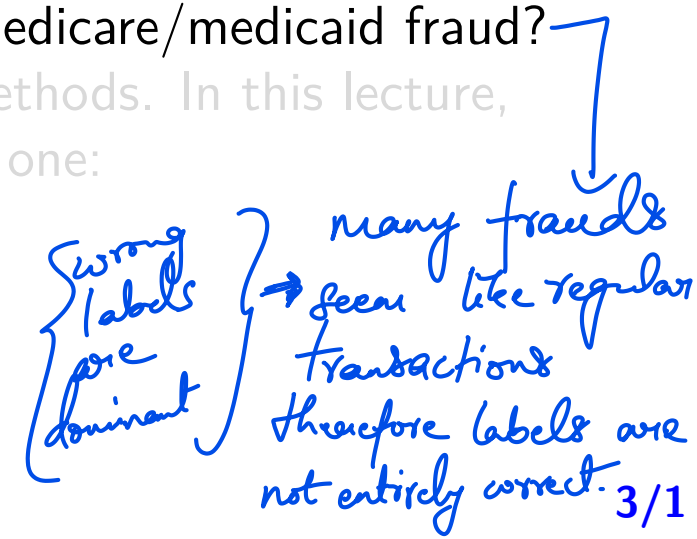
$\{1,\ldots,k\}$: multiclass classification.

- **Classification**: supervised learning when outcomes are categorical (a.k.a. qualitative)
- The categorical outcomes (responses) are usually called *class labels*.
- Both classification and regression are supervised learning
- Classification is perhaps the most widely used machine learning methods. Examples include email spam filter, credit card fraud detection, automatic cancer diagnosis, etc.
- In applications where (acurate) "labels" in theory exist, but we do not have access to them, we cannot formulate them as classification problems. Name one example? How about medicare/medicaid fraud?
- There are many off-the-shelf classification methods. In this lecture, we will begin with the most common (basic) one:
  - logistic regression

Classification

- **Classification**: supervised learning when outcomes are categorical (a.k.a. qualitative)
- The categorical outcomes (responses) are usually called *class labels*.
- Both classification and regression are supervised learning
- Classification is perhaps the most widely used machine learning methods. Examples include email spam filter, credit card fraud detection, automatic cancer diagnosis, etc.
- In applications where (acurate) "labels" in theory exist, but we do not have access to them, we cannot formulate them as classification problems. Name one example? How about medicare/medicaid fraud?
- There are many off-the-shelf classification methods. In this lecture, we will begin with the most common (basic) one:
  - logistic regression





Classification

- **Classification**: supervised learning when outcomes are categorical (a.k.a. qualitative)
- The categorical outcomes (responses) are usually called *class labels*.
- Both classification and regression are supervised learning
- Classification is perhaps the most widely used machine learning methods. Examples include email spam filter, credit card fraud detection, automatic cancer diagnosis, etc.
- In applications where (acurate) "labels" in theory exist, but we do not have access to them, we cannot formulate them as classification problems. Name one example? How about medicare/medicaid fraud?
- There are many off-the-shelf classification methods. In this lecture, we will begin with the most common (basic) one:
  - logistic regression

What is the usual objective for classification?

- Binary classification is the most common classification scenario
- Features $X \in R^p$ and class labels $Y \in \{0, 1\}$
- A classifier h is some function (usually data-dependent function) that maps the feature space into the label space. One can think of a classifier as a data-dependent partition of the feature space
- The classification error (risk) is the probability of misclassification. In other words:

  $P(h(X) \neq Y)$, where P is regarding the joint distribution of (X, Y).
- $P(h(X) \neq Y)$ is usually denoted by $R(h)$
- *Often* (NOT always), we construct classifiers to minimize the classification error.

- Note that the classification error can be decomposed into two parts

$$P(h(X) \neq Y) = P(h(X) \neq Y | Y = 0) \cdot P(Y = 0) + P(h(X) \neq Y | Y = 1) \cdot P(Y = 1)$$

we will talk more about this decomposition in future lectures

Type I error

Type II error

Let A & B be 2 events.

$P(A) = P(A \& B) + P(A \& B^c)$

$= P(A|B) \cdot P(B) + P(A|B^c) \cdot P(B^c)$

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

Q. $P(A \text{ and } B) = 0.6$, $P(B^c|A) = 0.4$, $P(B) = ?$

$P(B|A) = 1 - P(B^c|A) = 0.3$

$P(B|A) = P(A \text{ and } B) / P(A) = 0.3 \Rightarrow P(A) = \frac{0.6}{0.3} = 2$ which is > 1.

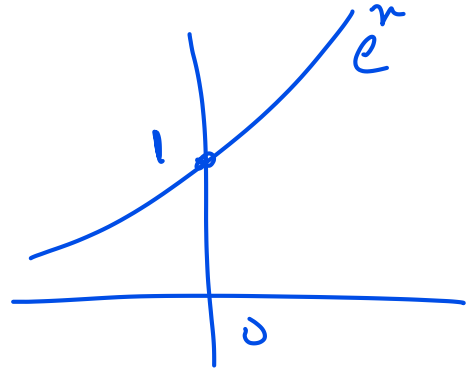
which means atleast one of the inputs are incorrect.

$\therefore$ 2 probabilities are incompatible

Why not linear regression?

- A general remark: before inventing new methods, we should ask why the existing ones do not suffice
- When the outcome variable has more than 2 categories. For example, an `income` variable has three levels: type A, type B, and type C
  - if we code *type* $A = 1$, *type* $B = 2$, *type* $C = 3$, and run linear regression, then
  - i). we have endorsed an ordering in the types
  - ii). we assumed the same difference between pairs
  - an equally reasonable coding *type* $C = 1$, *type* $B = 2$, *type* $A = 3$ will imply a totally different relationship among the three types
  - each of these codings will lead to different predictions
- When the outcome variable has 2 categories $\hat{y}_i \in (-\infty, \infty)$
  - we can introduce *dummy variables*
  - and cut the predicted y's at some level, i.e., declare prediction above that level of class 1, and 0 otherwise
  - but this approach is usually inferior to methods that specifically designed for classification

Logistic regression

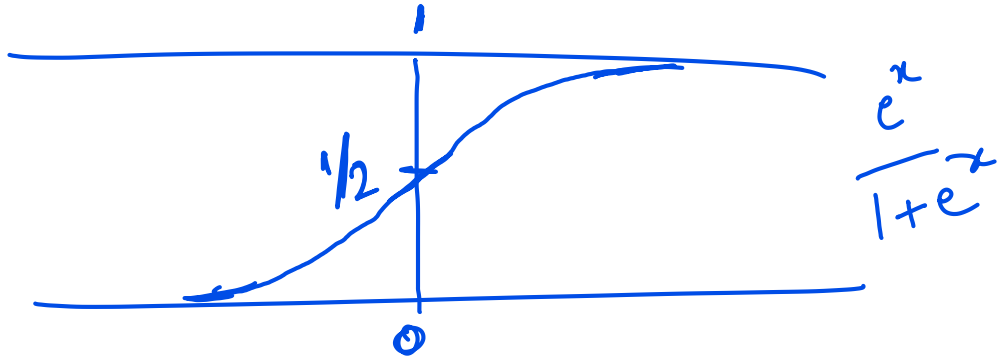


classification models this prob unlike regression

- Model the conditional probability $P(Y = 1|X = x)$ (compare with linear model)
- The logistic (a.k.a. sigmoid) function

$$0 \leq \quad f(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \quad \leq 1$$

- Logistic regression model: $P(Y = 1|X = x) = f(x)$. + ~~&~~ not required as uncertainty already taken care of.
- Plot the sigmoid function when $\beta_0 = 0$ and $\beta_1 = 1$



Logistic regression

can be deduced from $P(Y=1|X=x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$

- The sigmoid function f takes values between 0 and 1; perfect for modeling probability
- Under the logistic regression model, the *log-odds* or *logit* is linear in the input variable X:

$$\log\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right) = \beta_0 + \beta_1 x$$

- In some books, the above equation is the definition of logistic regression model, or called *logit model*. These two definitions are equivalent
- For logit function: https://en.wikipedia.org/wiki/Logit
- β_1 can be interpreted as the average change in log-odds associated with a one-unit increase in X
- β_1 does NOT correspond to the change in $P(Y=1|X=x)$ associated with 1 one-unit increase in X
- Q: recall the interpretation of the coefficients in linear regression, and find out the difference

Fitting Logistic regression

- The coefficients β_0 and β_1 in the sigmoid function are unknown
- Need to estimate them from training data
- Q: recall linear regression, what criterion did we use to find the coefficient estimates? Least Squares, in logistic, y modelling does not matter
- Given training data (pairs are independent of each other)

$$\{(x_1, y_1), \cdots, (x_n, y_n)\}$$

- And let $p(x) = P(Y = 1|X = x)$. We would like to find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that they maximize the *likelihood function* $l(\beta_0, \beta_1)$:

$$l(\beta_0, \beta_1) = \Pi_{i:y_i=1} p(x_i) \Pi_{i':y_{i'}=0} (1 - p(x_{i'}))$$

- This is called a *maximum likelihood approach*
- The least squares method for linear regression is in fact a maximum likelihood approach

About likelihood function

Frequentist : parameter has / exact value
Bayesian : parameter values keep varying

- Likelihood function is a *frequentist* idea
- To motivate the idea, suppose you know the distirbution is $\mathcal{N}(\mu, 1)$ with some known parameter μ, and you have observed a few points in the neighborhood of 0, which μ has the best opportunity to have produced these points?
- Q: Given observations (x, y pairs): $\{(2,1),(3,0)\}$, write down the **likelihood function** based on logistic regression model (Hint: start with only one pair $(2,1)$)

$$l(\beta_0,\beta_1) = P(Y=1 \mid X=2) \cdot P(Y=0 \mid x=3)$$

$$= \frac{e^{\beta_0+2\beta_1}}{1+e^{\beta_0+2\beta_1}} \cdot \frac{1}{1+e^{\beta_0+3\beta_1}}$$

Q1: about likelihood function

Given 3 paris of (x, y) observations $(2, 1)$, $(0, 1)$, and $(4, 0)$, write down the likelihood function based on logistic regression model

$$P(Y=1|X=2) \cdot P(Y=1|X=0) \cdot P(Y=0|X=4)$$

$$= \frac{e^{\beta_0+2\beta_1}}{1+e^{\beta_0+2\beta_1}} \cdot \frac{e^{\beta_0}}{1+e^{\beta_0}} \cdot \frac{1}{1+e^{\beta_0+4\beta_1}}$$

Q2: What's your respone to the following comment?

$$P(Y=1|x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

↳ Since this is probability, uncertainty is already modelled into it.

Your logistic regression model does not have that random error term ϵ, so it must be wrong if you want to model it as serious as a statistician.

Q3: Write down a logistic regression model with two independent variables

$$P(Y=1 \mid X_1 = x_1, X_2 = x_2) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}.$$

Please write down your answer without consulting any slides or books.

Q4:

$$\frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}}$$

plug in $x_1 = 40$, $x_2 = 3.5$ $\simeq 37.75\%$

Suppose we collect data for a group of students in a statistics class with variables X_1=hours studied, X_2= undergrad GPA, and Y=receive an A. We fit a logistic regression and produce estimated coefficients $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$. (Take $Y = 1$ to mean `receive an A`)

- Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.5 to get an A in the class.
- How many hours would the student in the previous part need to study to have a 50% chance of getting an A in the class?

$$\frac{e^{-6+0.05x_1+3.5}}{1+e^{-6+0.05x_1+3.5}} = 0.5$$

$$\Rightarrow e^{-6+0.05x_1+3.5} = 1$$

$$\Rightarrow -6+0.05x_1+3.5 = 0$$

$$0.05x_1 = 2.5$$

$$\Rightarrow x_1 = 50 \text{ hours}$$

Q5:

$$\Pi\left(\frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2 \cdots}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2 \cdots}} > c\right)$$

Π → indicator function.

probability of $y=1$ therefore $> c$ should indicate classification of 1.

Other than the python implementation, there is something really important that we haven't talked about. What is this missing piece in creating a classifier? threshold of classification.

Q6. Logistic regression has a linear decision boundary.

$$\Pi\left(\frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}} > c\right) \Rightarrow \text{decision boundary is precisely at } c.$$

default choice $c = 0.5$.

why though? $P(Y=1, X=x) > 0.5$ as $P(Y=1, X=x) > P(Y=0 \mid X=x)$

how?

$$e^{\beta_0+\beta_1 x} = c + c e^{\beta_0+\beta_1 x} \Rightarrow (1-c)\, e^{\beta_0+\beta_1 x+\ldots} = c$$

$$\Rightarrow e^{\beta_1+\beta_1 x} = c/(1-c)$$ as $0 < c < 1$ then $\frac{c}{1-c}$ is positive

$$\Rightarrow \beta_0+\beta_1 x_1+\beta_2 x_2 = \log\left(\frac{c}{1-c}\right) \Rightarrow \boxed{\beta_0 - \log\left(\frac{c}{1-c}\right) + \beta_1 x_1 + \beta_2 x_2 = 0}$$