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

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 "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 "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 "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

Why not linear regression?

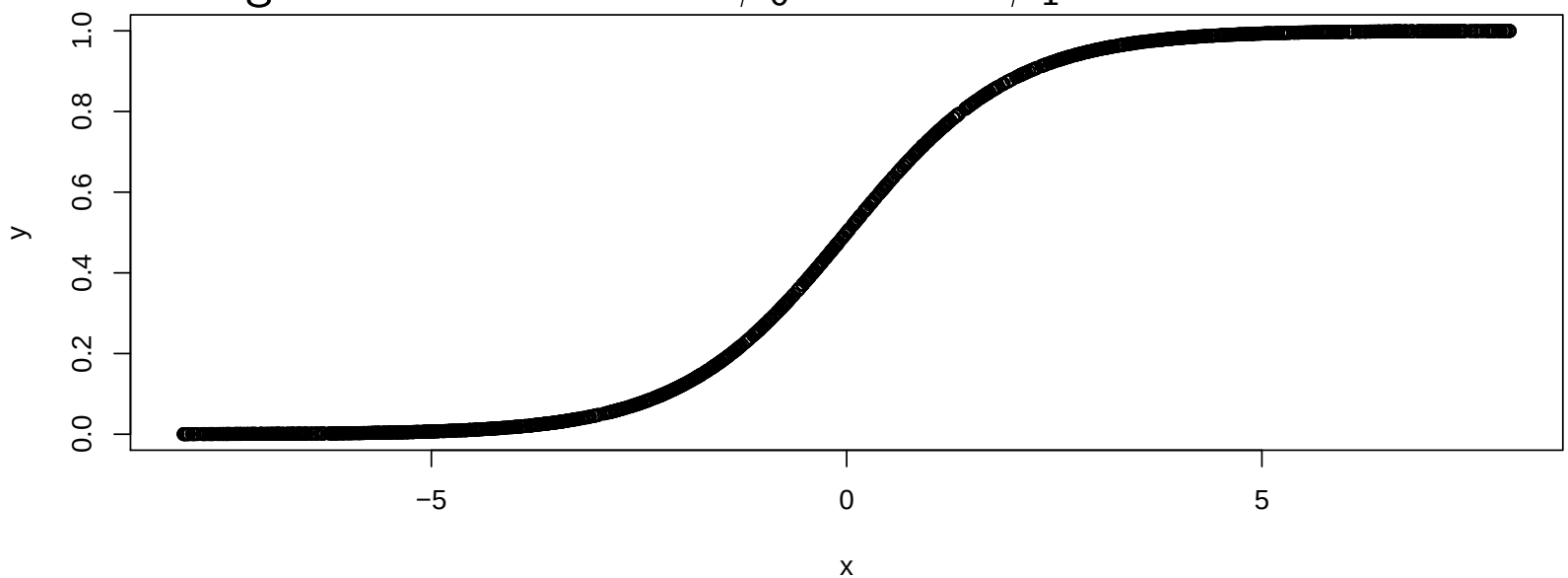
- A general remark: before inventing new methods, we should ask why the existing ones do not suffice
- When the outcome variable has more than 2 categories. For example, an `income` variable has three levels: type A, type B, and type C
  - if we code *type* $A = 1$, *type* $B = 2$, *type* $C = 3$, and 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
  - 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

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

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

- Logistic regression model: $P(Y = 1|X = x) = f(x)$.
- Plot the sigmoid function when $\beta_0 = 0$ and $\beta_1 = 1$



Logistic regression

- 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?
- 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

- 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 $(2, 1)$)

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

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

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

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

Q4:

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 $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

- 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?

Q5:

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?