CS 383: Machine Learning

Prof Adam Poliak

Fall 2024

10/08/2024

Lecture 13

Announcements

No lecture tomorrow Wednesday 10/09

HW04 naive Bayes due Wednesday 10/09

HW05 Logisitc Regression due Tuesday after Fall break

Midterm 1 on Thursday after fall break

Thursday reading quiz: https://see.stanford.edu/materials/aimlcs229/cs229-notes1.pdf - pages 16-19

Considerations when using a decision tree in medical applications

Advantages

- Very fast, gives a quick guide
- Could be more objective across different patient circumstances

Potential issues:

- Requires lots of historical data
- Decision trees work with how symptoms and other factors were historically measured
- If a new type of scan, xray, etc becomes available, we won't have historical data to incorporate into the algorithm
- Decision trees do not elegantly handle continuous features
- Patient may have a condition not seen in the training data
- 75% is not an optimal accuracy!

Outline

Logistic regression

Decision boundaries

Likelihood functions

Logistic regression cost function

SGD for logistic regression

Linear regression for classification

Case Study: you need to identify the medical condition of a patient in the emergency room on the basis of their symptoms.

Possible conditions (*y*) are:

- Stroke
- Drug overdose
- Epileptic seizure

1) If you were forced to use linear regression for this problem, how could you encode *y* to make it real-valued?

2) What issues arise with making y real-valued?

3) What if you just had two outcomes (i.e. stroke and drug overdose) -- why is linear regression still not a good choice?

Linear regression for classification

Case Study: you need to identify the medical condition of a patient in the emergency room on the basis of their symptoms.

Possible conditions (*y*) are:

- Stroke
- Drug overdose
- Epileptic seizure

1) If you were forced to use linear regression for this problem, how could you encode *y* to make it real-valued?

You could choose stroke=0, drug overdose=1, epileptic seizure=2 (or some permutation)

2) What issues arise with making y real-valued?

3) What if you just had two outcomes (i.e. stroke and drug overdose) -- why is linear regression still not a good choice?

Linear regression for classification

Case Study: you need to identify the medical condition of a patient in the emergency room on the basis of their symptoms.

Possible conditions (*y*) are:

- Stroke
- Drug overdose
- Epileptic seizure

1) If you were forced to use linear regression for this problem, how could you encode *y* to make it real-valued?

You could choose stroke=0, drug overdose=1, epileptic seizure=2 (or some permutation)

2) What issues arise with making y real-valued?

Assumes some *ordering* of the outcomes that is probably not there!

3) What if you just had two outcomes (i.e. stroke and drug overdose) -- why is linear regression still not a good choice?

Linear regression for classification

Case Study: you need to identify the medical condition of a patient in the emergency room on the basis of their symptoms.

Possible conditions (*y*) are:

- Stroke
- Drug overdose
- Epileptic seizure

1) If you were forced to use linear regression for this problem, how could you encode *y* to make it real-valued?

You could choose stroke=0, drug overdose=1, epileptic seizure=2 (or some permutation)

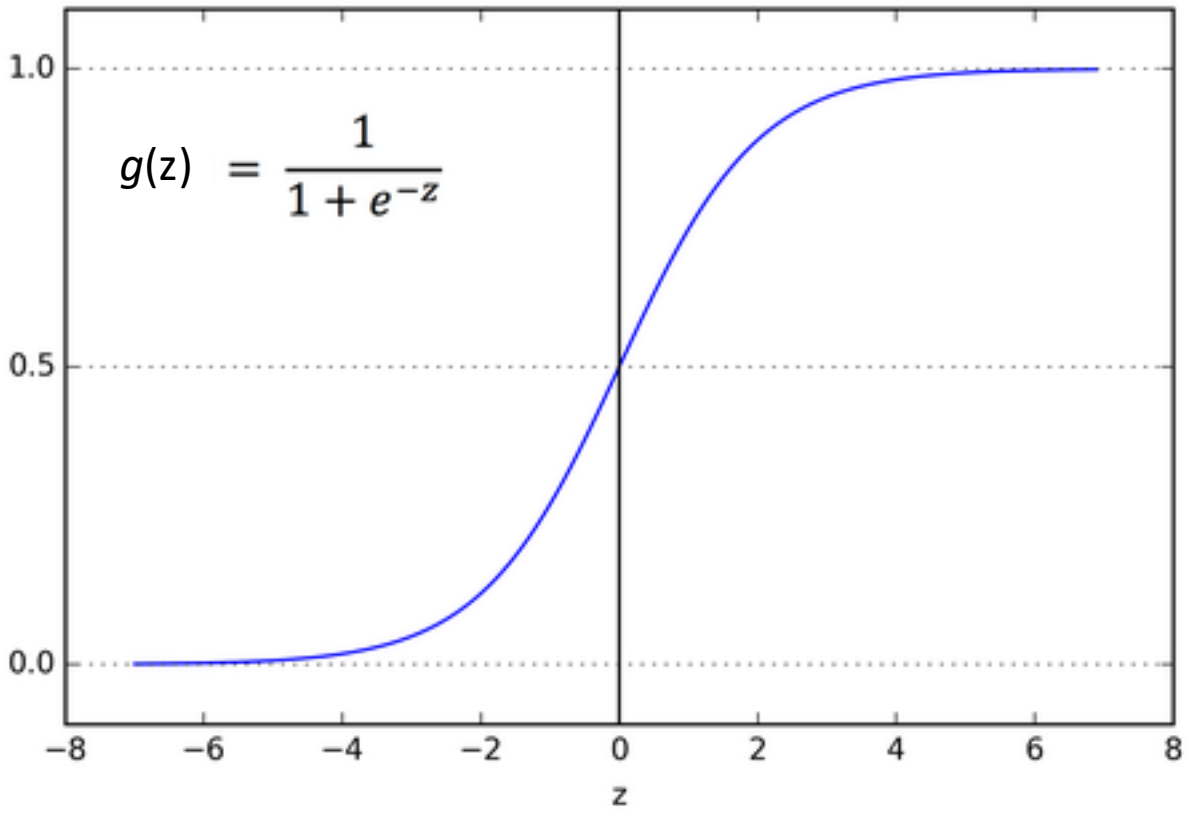
2) What issues arise with making y real-valued?

Assumes some *ordering* of the outcomes that is probably not there!

3) What if you just had two outcomes (i.e. stroke and drug overdose) -- why is linear regression still not a good choice?

The range of a linear function (i.e. *y* values) is [-∞, ∞], but we want [0, 1]

Logistic (sigmoid) function



Log-likelihood functions

Handout

Cost Function for Logistic Regression

SGD for Logistic Regression