Lec Mon

Monday, October 22, 2018 3:58 PM

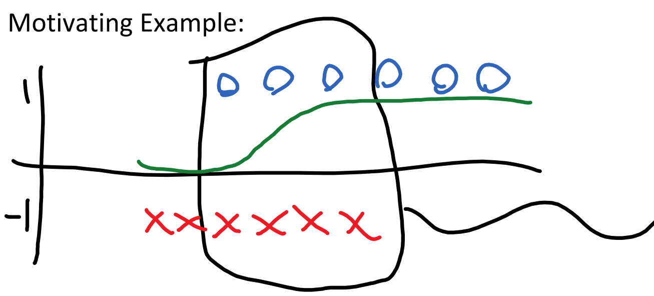
Recap: Perceptron
- learning a linear separator
- corrects itself on each failed prediction in training

Today: Logistic Regression
- a probabilistic model - we predict how likely the result is to take on one value over another

Online Learning vs. Batch Learning
- Online Learning takes one example at a time and uses them one by one to train the model
- Batch takes all the data and updates the model using multiple data at once
- Pros of Batch
  - takes less iterations
  - may be able to see some geometry or trend in all the data
- Pros of Online
  - model is more adaptable to new data
  - individual updates are usually simpler, resulting in smaller time complexity
  - space complexity smaller

Logistic Regression is Classification, but…
- output is discrete valued (-1 or 1) but we output a probability
- P(y=1 | x)
- as opposed to the perceptron algorithm which just gives a 1 or -1

Motivating Example:

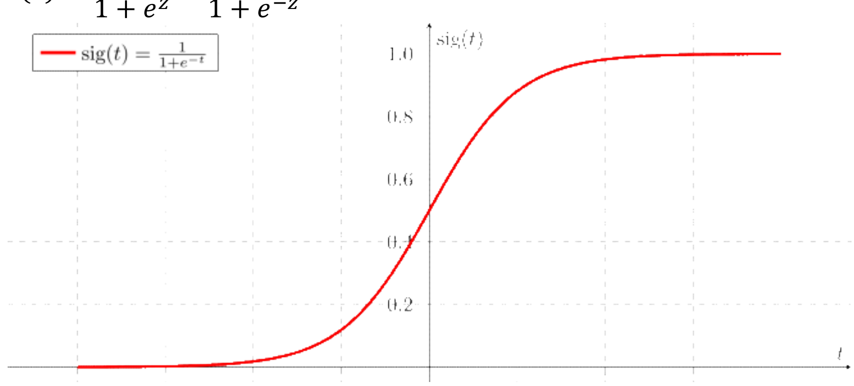


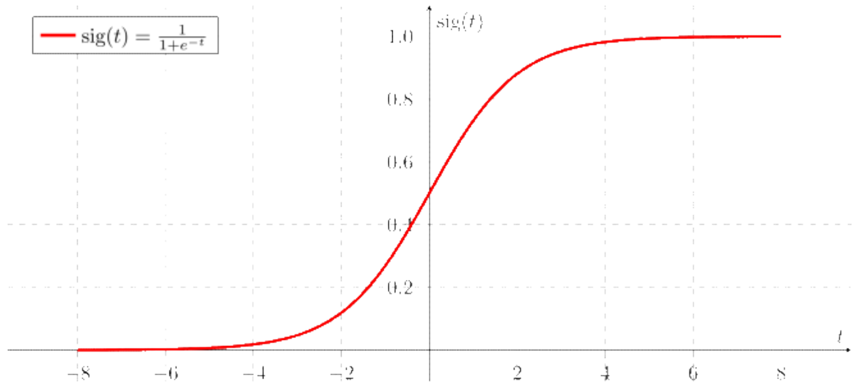
How to classify this region? A probability would be more descriptive

In Logistic Regression we use a logistic curve (green above) to assign probabilities

Sigmoid Function

$$\sigma(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$





What is $\frac{d}{dz}\frac{e^z}{1+e^z}$?

$$= \frac{-e^{-z}}{-(1+e^{-z})^2} = \frac{e^{-z}+1-1}{(1+e^{-z})^2} = \frac{1}{1+e^{-z}} - \frac{1}{(1+e^{-z})^2}$$

$$= \frac{1}{1+e^{-z}}\left(1 - \frac{1}{1+e^{-z}}\right) = \sigma(z)(1-\sigma(z))$$

How to make a prediction based on the probability?
if P(y=1|x) > 0.5, we predict 1. otherwise, -1

In multiple dimensions, $P(y=1|\boldsymbol{x};\boldsymbol{w}) = \sigma(\boldsymbol{w}^T\boldsymbol{x}) = \frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}}}$

When does y=1?

$\frac{1}{1+e^{-w^Tx}} \geq 0.5$

$1+e^{-w^Tx} \leq 2$

$e^{-w^Tx} \leq 1$

$w^Tx \geq 0$ (just like perceptron!)

The decision boundary is linear

We draw papers from an envelope. They're either purple or yellow. Let's say we drew k yellow, n-k purple. What is the probability that maximizes the possibility of this outcome?

$\max\left(\theta^k(1-\theta^k)\right)$

turns out it's just k/n which I probably could've told you in middle school