Logistic regression

DSE 220

Outline

1. Conditional probability estimation for binary labels
2. Learning a logistic regression model
3. Logistic regression in use

Uncertainty in prediction

Can we usually expect to get a perfect classifier, if we have enough training data?

Problem 1: Inherent uncertainty

The available features x do not contain enough information to perfectly predict y, e.g.,

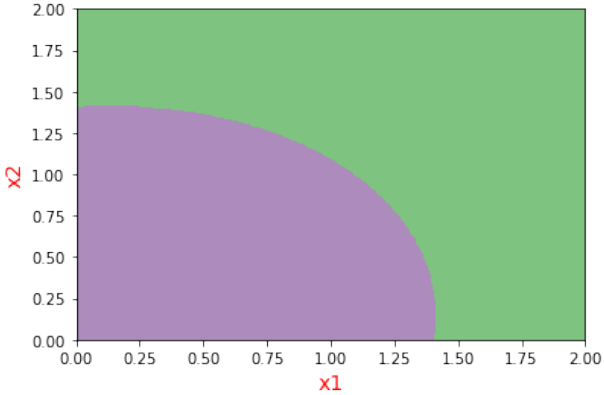
- x = complete medical record for a patient at risk for a disease
- y = will he/she contract the disease in the next 5 years?

Uncertainty in prediction, cont'd

Can we usually expect to get a perfect classifier, if we have enough training data?

Problem 2: Limitations of the model class

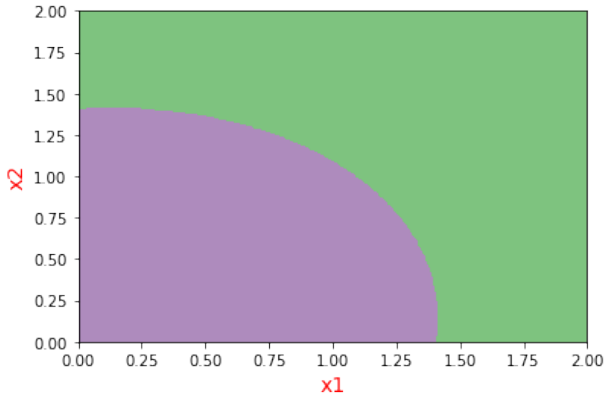
The type of classifier being used does not capture the decision boundary, e.g. using linear classifiers with:



Conditional probability estimation for binary labels

- Given: data set of pairs (x, y) with $x \in \mathbb{R}^d$ and $y \in \{-1, 1\}$
- Return a classifier that also gives probabilities $\Pr(y = 1|x)$

Simplest case: using a linear function of x.



A linear model for conditional probability estimation

For data $x \in \mathbb{R}^d$, classify and return probabilities using a linear function

$$w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b \;=\; w \cdot x + b$$

where $w = (w_1, \ldots, w_d)$.

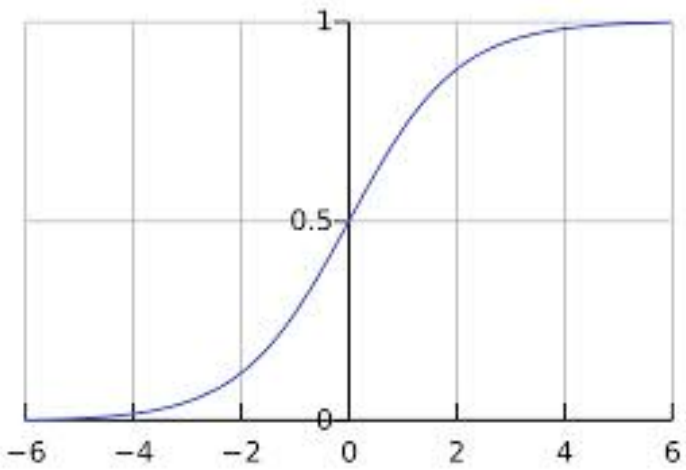
The probability of $y = 1$:

- Increases as the linear function grows.
- Is 50% when this linear function is zero.

How can we convert $w \cdot x + b$ into a probability?

The squashing function

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



The logistic regression model

Binary labels $y \in \{-1, 1\}$. Model:

$$\Pr(y = 1|x) = \frac{1}{1 + e^{-(w \cdot x + b)}}$$

What is $\Pr(y = -1|x)$?

Summary: logistic regression for binary labels

- Data $x \in \mathbb{R}^d$
- Binary labels $y \in \{-1, 1\}$

Model parametrized by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$:

$$\Pr_{w,b}(y|x) = \frac{1}{1 + e^{-y(w \cdot x + b)}}$$

Learn parameters w, b from data

Outline

1. Conditional probability estimation for binary labels
2. Learning a logistic regression model
3. Logistic regression in use

The learning problem

Given data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, 1\}$

Maximum-likelihood: pick $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ that maximize

$$\prod_{i=1}^{n} \Pr_{w,b}(y^{(i)} \mid x^{(i)})$$

Take log to get **loss function**

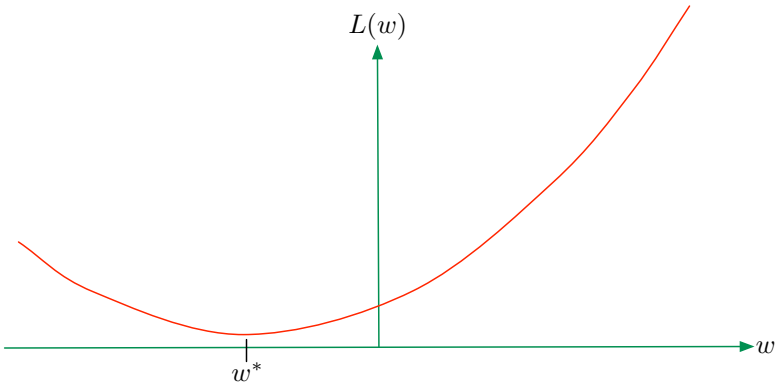
$$L(w, b) \;=\; -\sum_{i=1}^{n} \ln \Pr_{w,b}(y^{(i)} \mid x^{(i)}) \;=\; \sum_{i=1}^{n} \ln(1 + e^{-y^{(i)}(w \cdot x^{(i)} + b)})$$

Goal: minimize $L(w, b)$.

As with linear regression, can absorb b into w.
Yields simplified loss function $L(w)$.

Convexity

- Bad news: no closed-form solution for w
- Good news: $L(w)$ is **convex** in w



How to find the minimum of a convex function? By **local search.**

Gradient descent procedure for logistic regression

Given $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, 1\}$, find

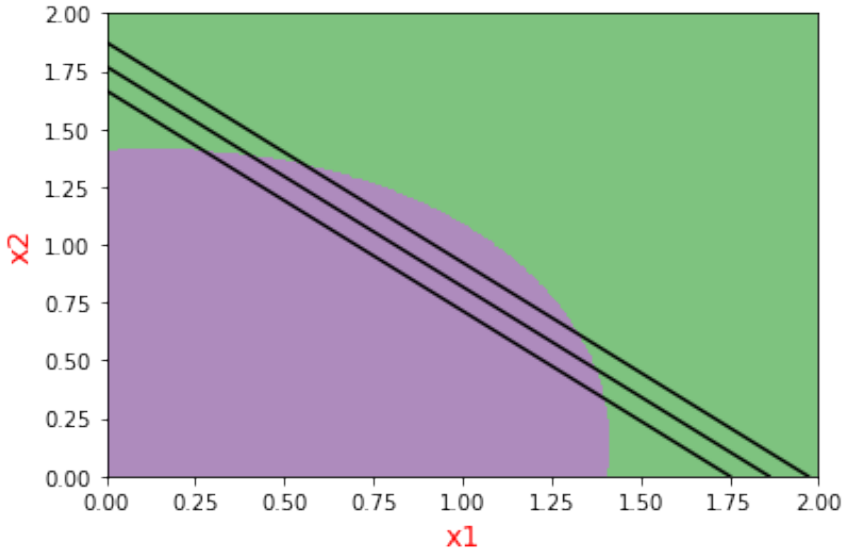
$$\arg\min_{w \in \mathbb{R}^d} L(w) \;=\; \sum_{i=1}^{n} \ln(1 + e^{-y^{(i)}(w \cdot x^{(i)})})$$

- Set $w_0 = 0$
- For $t = 0, 1, 2, \ldots$, until convergence:

$$w_{t+1} \;=\; w_t \;+\; \eta_t \sum_{i=1}^{n} y^{(i)} x^{(i)} \underbrace{\Pr_{w_t}(-y^{(i)} | x^{(i)})}_{\text{doubt}_t(x^{(i)}, y^{(i)})},$$

where η_t is a "step size"

Toy example



Outline

1. Conditional probability estimation for binary labels
2. Learning a logistic regression model
3. Logistic regression in use

Example: Sentiment data

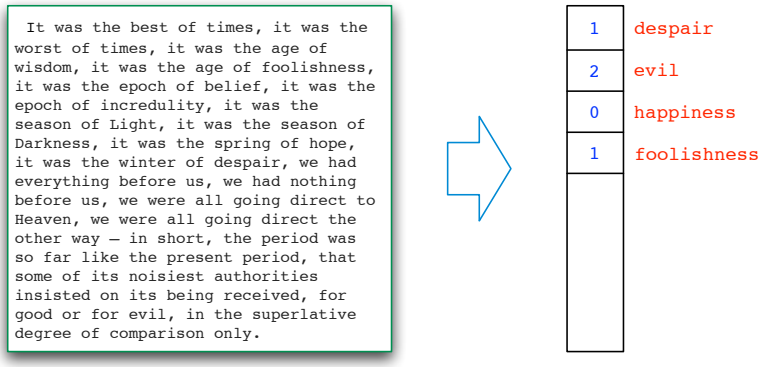
Data set: sentences from reviews on Amazon, Yelp, IMDB.
Each labeled as positive or negative.

- `Needless to say, I wasted my money.`
- `He was very impressed when going from the original battery to the extended battery.`
- `I have to jiggle the plug to get it to line up right to get decent volume.`
- `Will order from them again!`

2500 training sentences, 500 test sentences

Handling text data

Bag-of-words: vectorial representation of text sentences (or documents).



- Fix V = some vocabulary.
- Treat each sentence (or document) as a vector of length $|V|$:

$$x = (x_1, x_2, \ldots, x_{|V|}),$$

where x_i = # of times the ith word appears in the sentence.

A logistic regression approach

Code positive as $+1$ and negative as -1.

$$\Pr_{w,b}(y \mid x) = \frac{1}{1 + e^{-y(w \cdot x + b)}}$$

Given $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, 1\}$, loss function

$$L(w, b) = \sum_{i=1}^{n} \ln(1 + e^{-y^{(i)}(w \cdot x^{(i)} + b)})$$

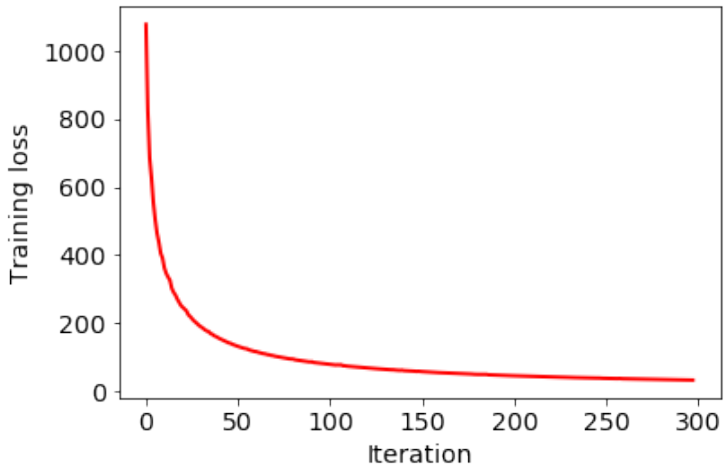
Convex problem with many solution methods, e.g.

- gradient descent, stochastic gradient descent
- Newton-Raphson, quasi-Newton

All converge to the optimal solution.

Local search in progress

Look at how loss function $L(w, b)$ changes over iterations of stochastic gradient descent.



Final model: **test error** 0.21.

Some of the mistakes

Not much dialogue, not much music, the whole film was shot as elaborately and aesthetically like a sculpture. 1

This film highlights the fundamental flaws of the legal process, that it's not about discovering guilt or innocence, but rather, is about who presents better in court. 1

You need two hands to operate the screen. This software interface is decade old and cannot compete with new software designs. -1

The last 15 minutes of movie are also not bad as well. 1

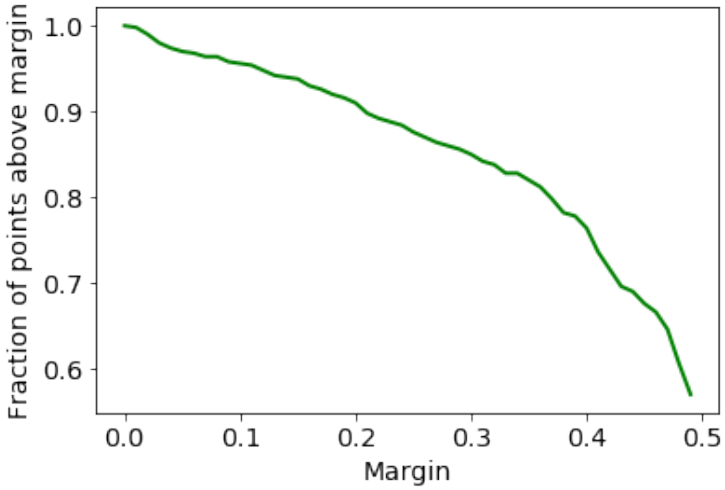
If you plan to use this in a car forget about it. -1

If you look for authentic Thai food, go else where. -1

Waste your money on this game. 1

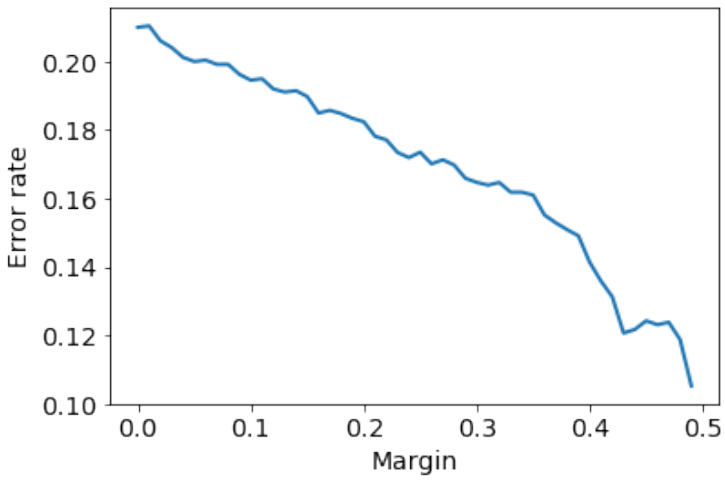
Margin and test error

$$\text{Margin on test pt } x = \left|\Pr_{w,b}(y = 1|x) - \frac{1}{2}\right| \quad .$$



Margin and test error

$$\text{Margin on test pt } x = \left| \Pr_{w,b}(y = 1|x) - \frac{1}{2} \right| \quad .$$



Interpreting the model

Words with the most positive coefficients

'sturdy', 'able', 'happy', 'disappoint', 'perfectly', 'remarkable', 'animation', 'recommendation', 'best', 'funny', 'restaurant', 'job', 'overly', 'cute', 'good', 'rocks', 'believable', 'brilliant', 'prompt', 'interesting', 'skimp', 'definitely', 'comfortable', 'amazing', 'tasty', 'wonderful', 'excellent', 'pleased', 'beautiful', 'fantastic', 'delicious', 'watch', 'soundtrack', 'predictable', 'nice', 'awesome', 'perfect', 'works', 'loved', 'enjoyed', 'love', 'great', 'happier', 'properly', 'liked', 'fun', 'screamy', 'masculine'

Words with the most negative coefficients

'disappointment', 'sucked', 'poor', 'aren', 'not', 'doesn', 'worst', 'average', 'garbage', 'bit', 'looking', 'avoid', 'roasted', 'broke', 'starter', 'disappointing', 'dont', 'waste', 'figure', 'why', 'sucks', 'slow', 'none', 'directing', 'stupid', 'lazy', 'unrecommended', 'unreliable', 'missing', 'awful', 'mad', 'hours', 'dirty', 'didn', 'probably', 'lame', 'sorry', 'horrible', 'fails', 'unfortunately', 'barking', 'bad', 'return', 'issues', 'rating', 'started', 'then', 'nothing', 'fair', 'pay'