Logistic regression

DSE 220

Classification

Conditional probability estimation

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?

In general, no.

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?

In such cases, we can make a prediction, but we'd also like to quantify the uncertainty in the prediction.

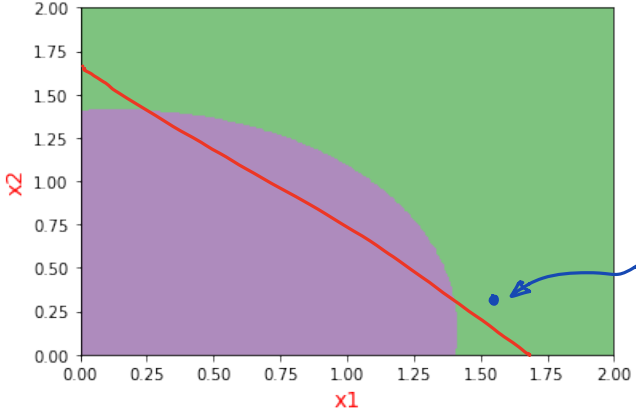
E.g. "there is an 80% chance the person will get the disease!"

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 $\Pr(y = 1|x)$

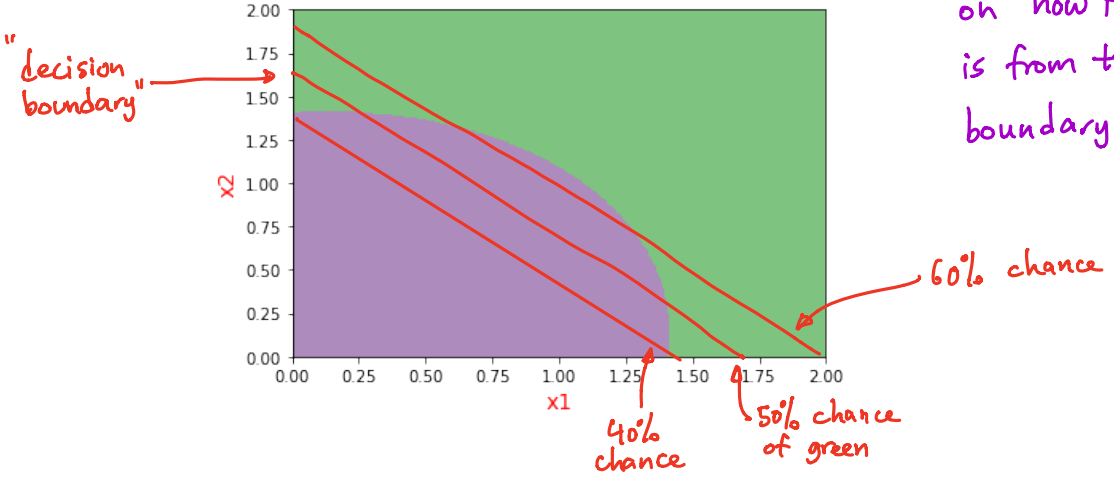
Conditional probability estimation for binary labels

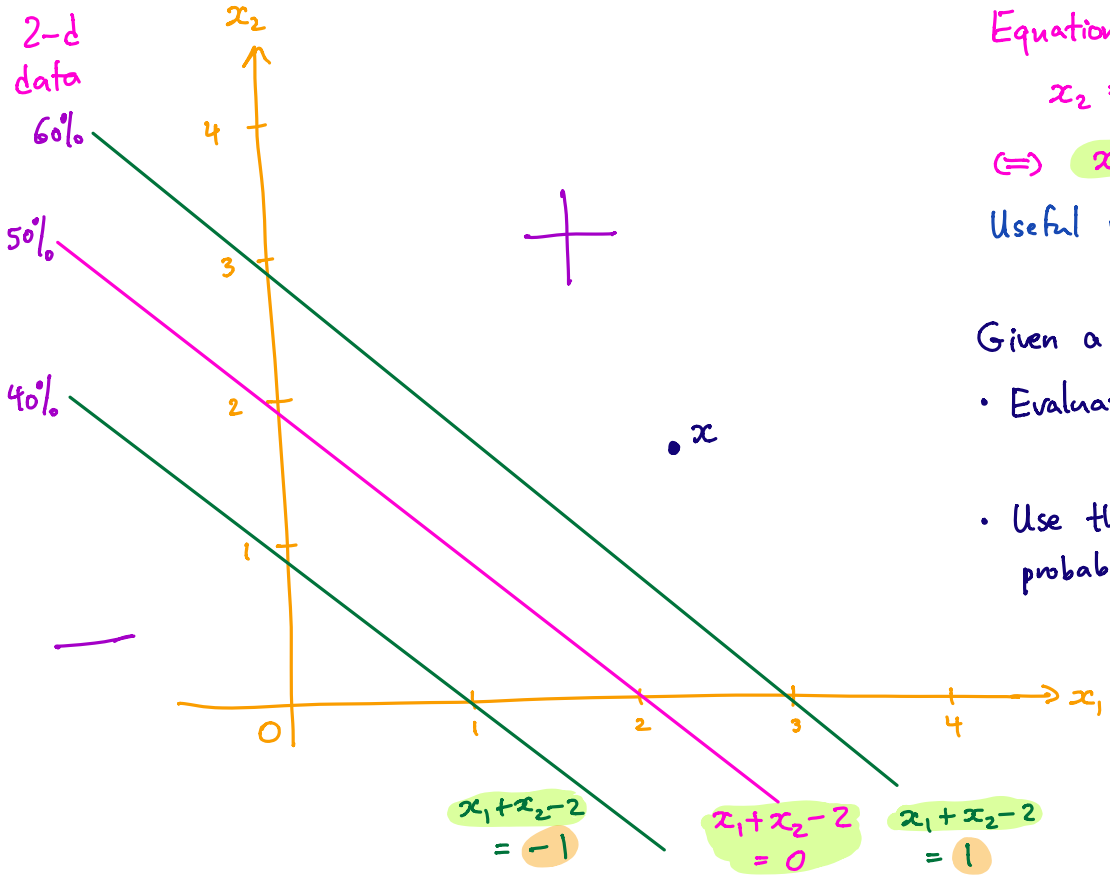
our convention. (other choices, like {0,1} or {1,2}, would be fine too)

- 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.

assign probabilities based on how far the point is from the decision boundary





Equation of decision boundary:

$$x_2 = -x_1 + 2$$

$$\Leftrightarrow \quad x_1 + x_2 - 2 = 0$$

Useful form: (linear fn) = 0

Given a point $x = (x_1, x_2)$:

- Evaluate the linear fn

  $$x_1 + x_2 - 2$$

- Use this value to predict the probability that $y = +1$
  - If value = 0: prob = 0.5
  - If value > 0: prob > 0.5
  - If value < 0: prob < 0.5

A linear model for conditional probability estimation

$x = (x_1, x_2, \ldots, x_d)$

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

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

linear function for data in d dimensions

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.

decision boundary: $w \cdot x + b = 0$

$w \cdot x + b > 0$: predict +1

the higher the value, the greater the probability

When $d = 2$: line

When $d = 3$: plane

When $d > 3$: hyperplane

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

some real number

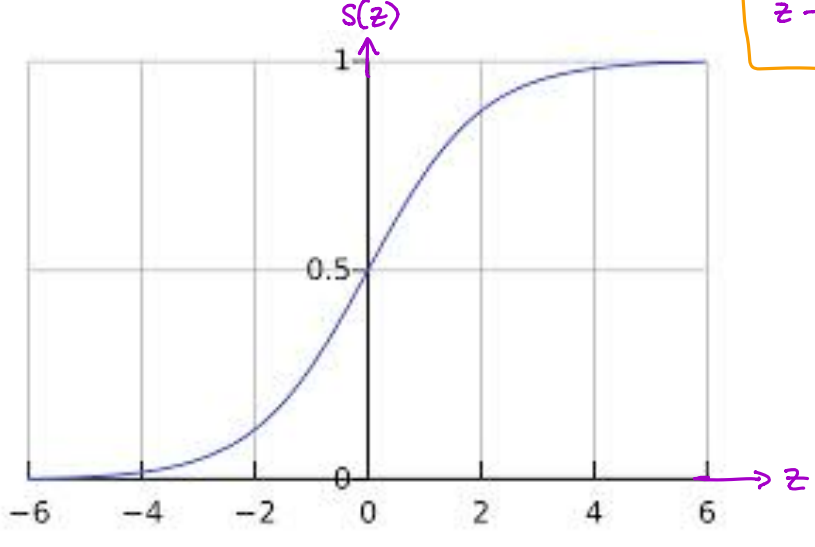
in the range $[0, 1]$

The squashing function

$e = 2.718\ldots$

$e^{-z} = \exp(-z)$

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



Here $z = w \cdot x + b$

$z = 0 \Rightarrow s(z) = 0.5$ (decision boundary)

$z \to \infty \Rightarrow s(z) \to 1$

$z \to -\infty \Rightarrow s(z) \to 0$

The logistic regression model

specified by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ [need to learn these using data]

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

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

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

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

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

Concise form that captures both cases, $y=-1$ and $y=1$:

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

works for both $y=+1$ and $y=-1$

Summary: logistic regression for binary labels

Worksheet 6 # 1, 2

- 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

log:

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

Maximum likelihood

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

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

$$\Pr(y^{(1)} \mid x^{(1)}) \times \Pr(y^{(2)} \mid x^{(2)}) \times \Pr(y^{(3)} \mid x^{(3)}) \times \cdots \times \Pr(y^{(n)} \mid x^{(n)})$$

Pick the parameters w,b that maximize the probability of the data.

① Maximizing $F(w,b)$ is the same as maximizing $\log F(w,b)$

② Maximizing $F(w,b)$ is the same as minimizing $-F(w,b)$

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

$$\log \prod_{i=1}^{n} \Pr(y^{(i)} | x^{(i)}) = \sum_{i=1}^{n} \log \Pr(y^{(i)} | x^{(i)})$$

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

Take log to get **loss function**

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

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

Loss function $L(w, b)$

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

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**

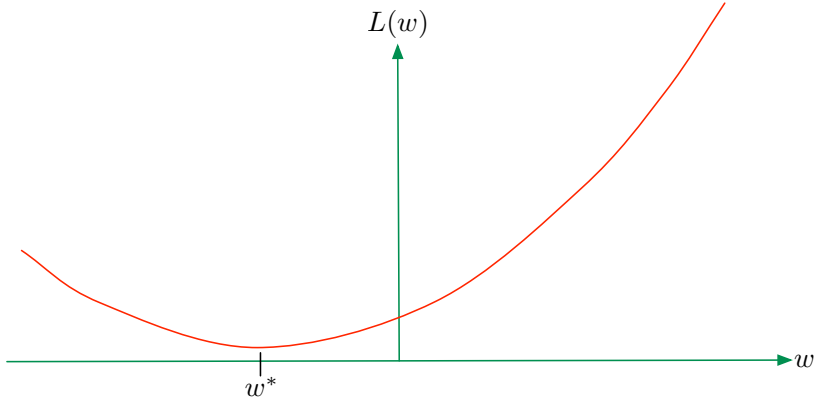
$$L(w, b) \;=\; -\sum_{i=1}^{n} \ln \Pr_{w,b}(y^{(i)} \mid x^{(i)}) \;=\; \sum_{i=1}^{n} \ln(1 + e^{-y^{(i)}(w \cdot x^{(i)} + b)})$$

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

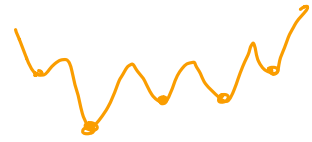
As with linear regression, can absorb b into w. (add an extra feature to x with constant value 1)
Yields simplified loss function $L(w)$.

Convexity

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

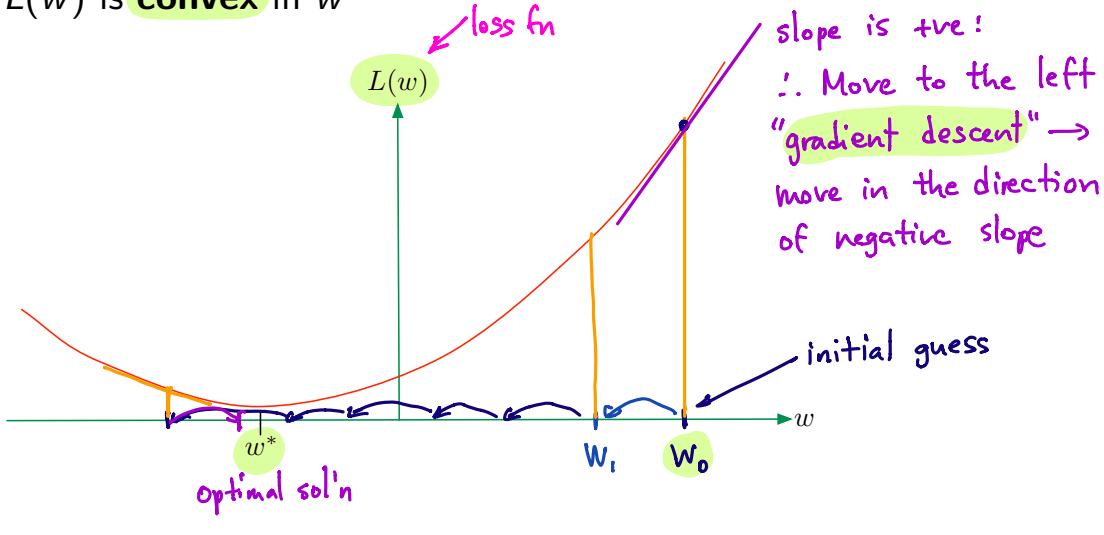


Convexity



"bowl-shaped"

- 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 $w_0 = 0$ ← vector of all-zeros
- For $t = 0, 1, 2, \ldots$, until convergence:

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

negative of the slope (gradient)

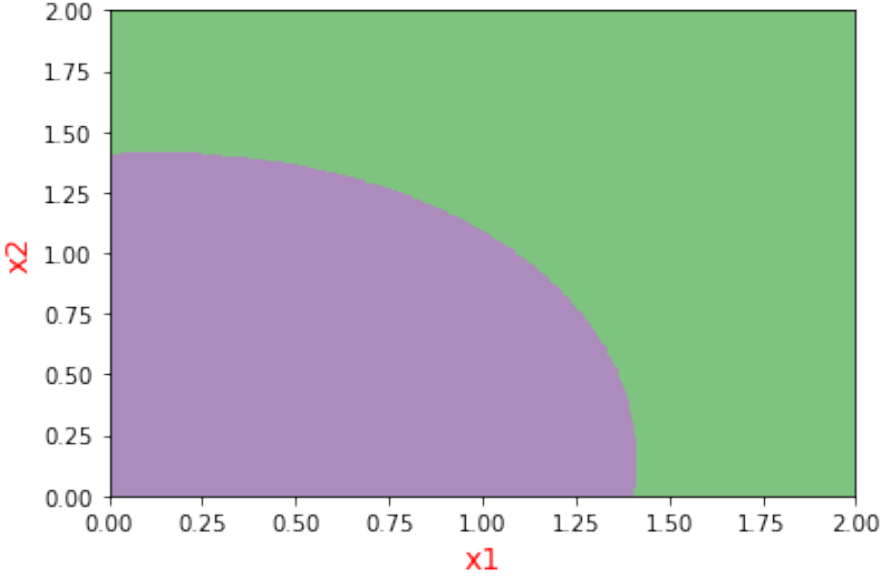
w_{t+1}: d-vector

η_t: step size

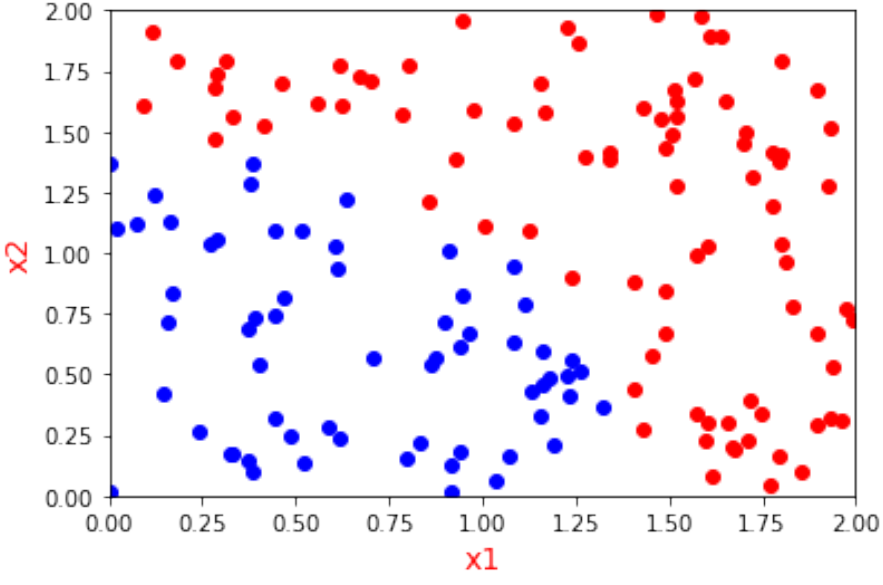
where η_t is a "step size"

add up all the $y^{(i)}x^{(i)}$, but each is weighted by the probability that the current w_t assigns to the WRONG label

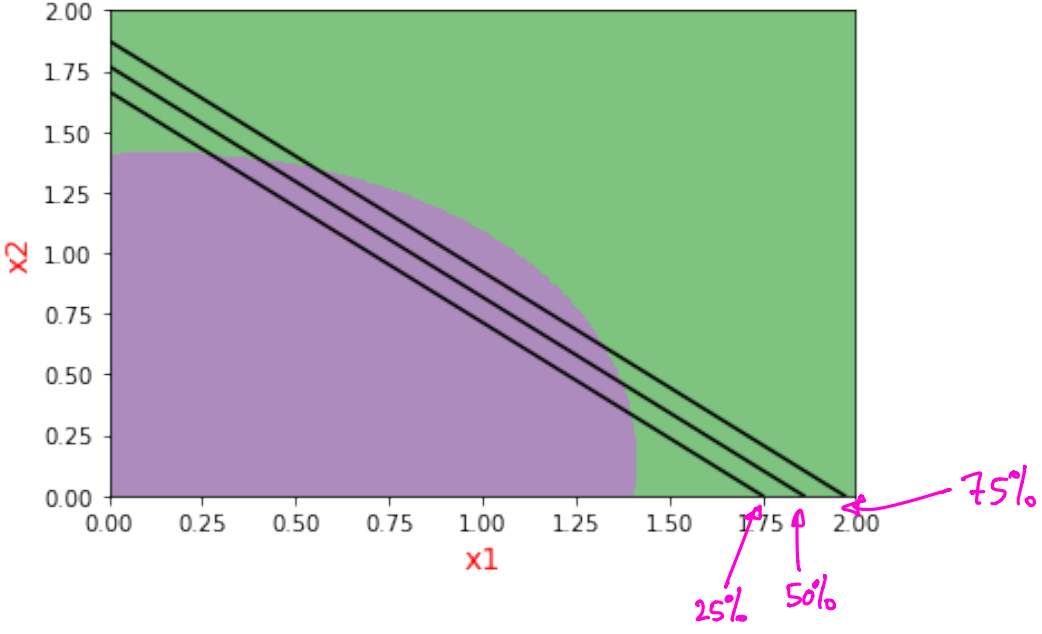
Toy example


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
x2
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
x1

Toy example


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
x2
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
x1

Toy example


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
x2
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
x1
75%
25%
50%

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. -1
- He was very impressed when going from the original battery to the extended battery. +1
- I have to jiggle the plug to get it to line up right to get decent volume. -1
- Will order from them again! +1

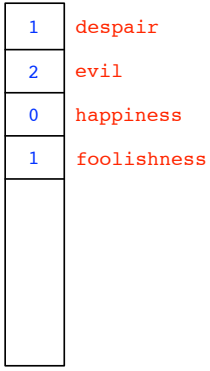
2500 training sentences, 500 test sentences Not much data!

To use LR: convert sentences into vectors

Handling text data

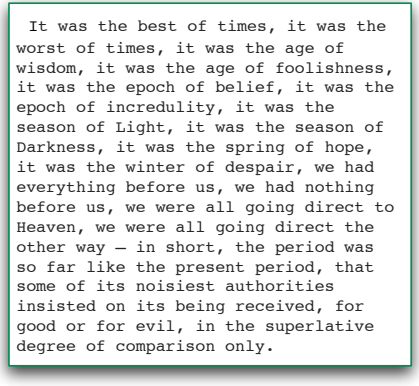
Bag-of-words: vectorial representation of text sentences (or documents).

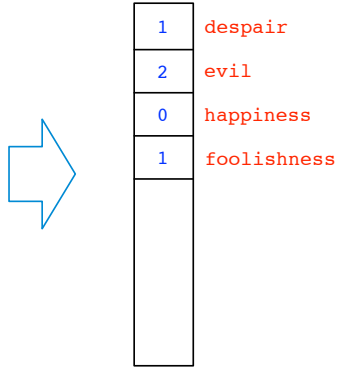
It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way – in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.



Handling text data

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





Each sentence becomes a 5000-dim vector of word counts.

Sparse

V = 5000 most common words in the reviews

- 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

5000-dim vector

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

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

$n = 2500$
(size of training set)

Learn w, b from this data.

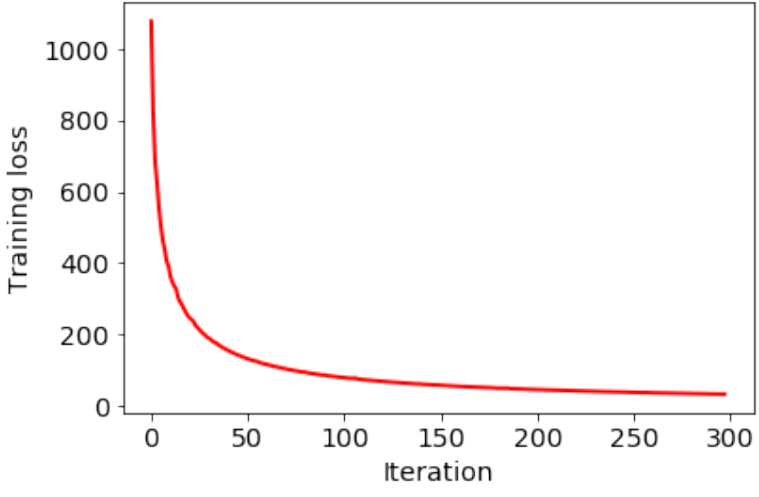
Convex problem with many solution methods, e.g.

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

All converge to the optimal solution.

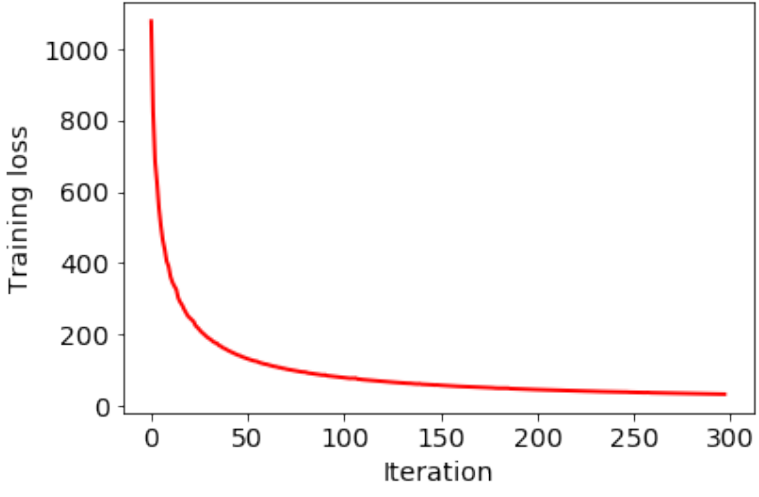
Local search in progress

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



Local search in progress

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



Loss kept going down, converged after 300 iterations.

Final model: **test error** 0.21.

Not too bad given — difficulty of task
— shortage of training data

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

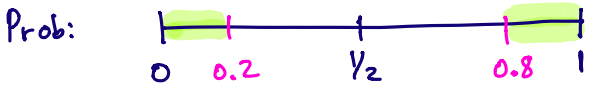
If you look for authentic Thai food, go else where. -1 ← true labels

Waste your money on this game. 1

Margin and test error

How far is the prob from 1/2?

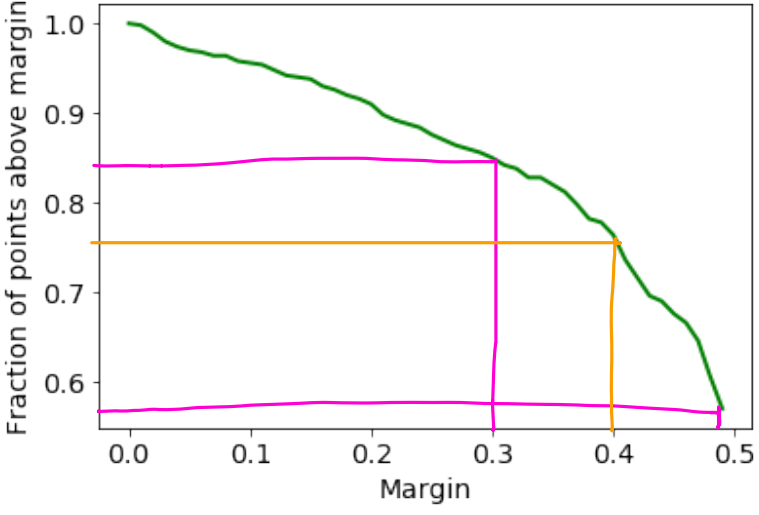
Margin on test pt $x = \left|\Pr_{w,b}(y = 1|x) - \frac{1}{2}\right|$



margin ≥ 0.3
⇔ prob lies in [0, 0.2]
or [0.8, 1]

About 83% of the test points have margin ≥ 0.3

About 55% of the test points have margin ≥ 0.49, ie. probs in the range [0, 0.01]
or [0.99, 1]

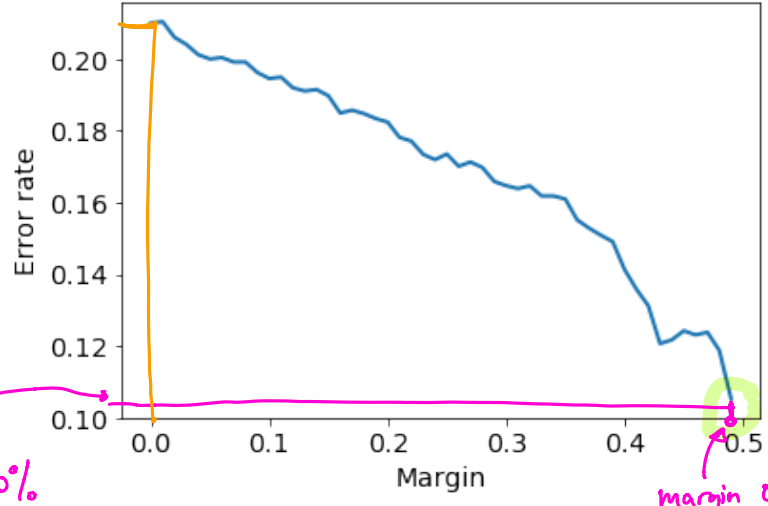


Despite a shortage of training data, this classifier is very confident!

Need to take these probabilities with a grain of salt.

Margin and test error

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



In this example, higher confidence (margin) ⇒ lower error rate

① Conditional probs are informative

② But the model is grossly overconfident

error rate on those points is ≈ 10%

probs in [0.99,1] or [0, 0.01]

margin 0.49 (we've seen that 55% of test data has this)

Interpreting the model

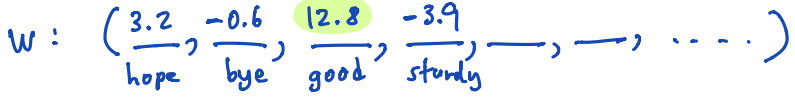
Vector w has an entry for every one of the 5000 vocab. words. Which words are the most important for prediction, ie. with the largest coefficients?

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'



Technology used in LR: loss fn, convexity, stochastic gradient descent, ...