An introduction to linear regression

Topics we'll cover

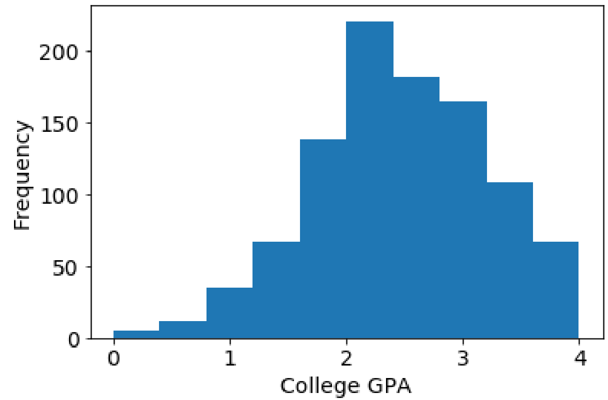
1. The regression problem in one dimension
2. Predictor and response variables
3. A loss function formulation
4. Deriving the optimal solution

Linear regression

Fitting a line to a bunch of points.

Example: college GPAs

Distribution of GPAs of students at a certain Ivy League university.

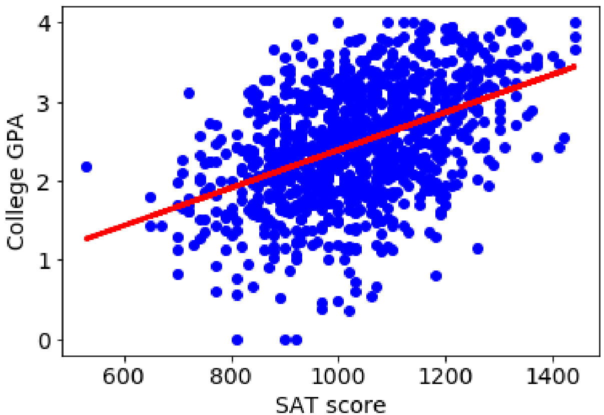


What GPA to predict for a random student from this group?

- Without further information, predict the **mean**, 2.47.
- What is the average squared error of this prediction?
  That is, $\mathbb{E}[((\text{student's GPA}) - (\text{predicted GPA}))^2]$?
  The **variance** of the distribution, 0.55.

Better predictions with more information

We also have SAT scores of all students.



Mean squared error (MSE) drops to 0.43.

This is a **regression** problem with:

- **Predictor variable**: SAT score
- **Response variable**: College GPA

Parametrizing a line

A line can be parameterized as $y = ax + b$ (a: slope, b: intercept).

The line fitting problem

Pick a line (parameters a, b) suited to the data, $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R} \times \mathbb{R}$

- $x^{(i)}, y^{(i)}$ are predictor and response variables, e.g. SAT score, GPA of ith student.
- Minimize the mean squared error,

$$\mathrm{MSE}(a, b) = \frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - (ax^{(i)} + b))^2.$$

This is the **loss function**.

Minimizing the loss function

Given $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$, minimize

$$L(a, b) = \sum_{i=1}^{n} (y^{(i)} - (ax^{(i)} + b))^2.$$

Linear regression

Topics we'll cover

1. Regression with multiple predictor variables
2. Least-squares regression
3. The least-squares solution

Diabetes study

Data from $n = 442$ diabetes patients.

For each patient:

- 10 features $x = (x_1, \ldots, x_{10})$

  age, sex, body mass index, average blood pressure, and six blood serum measurements.

- A real value y: the progression of the disease a year later.

Regression problem:

- **response** $y \in \mathbb{R}$
- **predictor variables** $x \in \mathbb{R}^{10}$

Least-squares regression

Linear function of 10 variables: for $x \in \mathbb{R}^{10}$,

$$f(x) \;=\; w_1 x_1 + w_2 x_2 + \cdots + w_{10} x_{10} + b \;=\; w \cdot x + b$$

where $w = (w_1, w_2, \ldots, w_{10})$.

Penalize error using **squared loss** $(y - (w \cdot x + b))^2$.

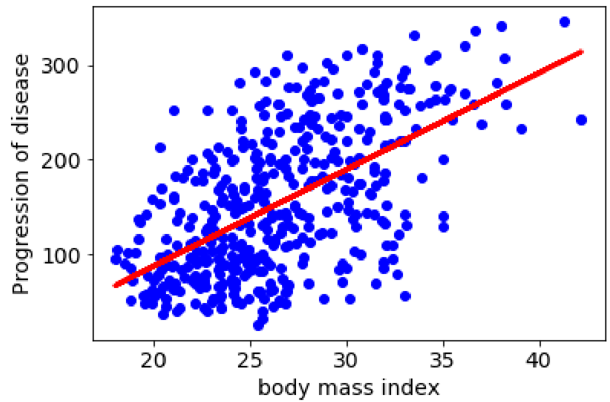
Least-squares regression:

- *Given:* data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \mathbb{R}$
- *Return:* linear function given by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$
- *Goal:* minimize the **loss function**

$$L(w, b) = \sum_{i=1}^{n} (y^{(i)} - (w \cdot x^{(i)} + b))^2 \qquad .$$

Back to the diabetes data

- No predictor variables: mean squared error (MSE) = 5930
- One predictor ('bmi'): MSE = 3890



- Two predictors ('bmi', 'serum5'): MSE = 3205
- All ten predictors: MSE = 2860

Least-squares solution 1

Linear function of d variables given by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$:

$$f(x) \;=\; w_1x_1 + w_2x_2 + \cdots + w_dx_d + b \;=\; w \cdot x + b$$

Assimilate the intercept b into w:

- Add a new feature that is identically 1: let $\widetilde{x} = (1, x) \in \mathbb{R}^{d+1}$

$$\begin{pmatrix} 4 & 0 & 2 & \cdots & 3 \end{pmatrix} \Longrightarrow \begin{pmatrix} 1 & 4 & 0 & 2 & \cdots & 3 \end{pmatrix}$$

- Set $\widetilde{w} = (b, w) \in \mathbb{R}^{d+1}$
- Then $f(x) = w \cdot x + b = \widetilde{w} \cdot \widetilde{x}$

Goal: find $\widetilde{w} \in \mathbb{R}^{d+1}$ that minimizes

$$L(\widetilde{w}) \;=\; \sum_{i=1}^{n} (y^{(i)} - \widetilde{w} \cdot \widetilde{x}^{(i)})^2$$

Least-squares solution 2

Write

$$X = \begin{pmatrix} \longleftarrow & \widetilde{x}^{(1)} & \longrightarrow \\ \longleftarrow & \widetilde{x}^{(2)} & \longrightarrow \\ & \vdots & \\ \longleftarrow & \widetilde{x}^{(n)} & \longrightarrow \end{pmatrix}, \quad y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{pmatrix}$$

Then the loss function is

$$L(\widetilde{w}) = \sum_{i=1}^{n} (y^{(i)} - \widetilde{w} \cdot \widetilde{x}^{(i)})^2 = \|y - X\widetilde{w}\|^2$$

and it minimized at $\widetilde{w} = (X^T X)^{-1}(X^T y)$.

Regularized linear regression

Topics we'll cover

1. Generalization
2. Regularization
3. Ridge regression
4. Lasso

Least-squares regression

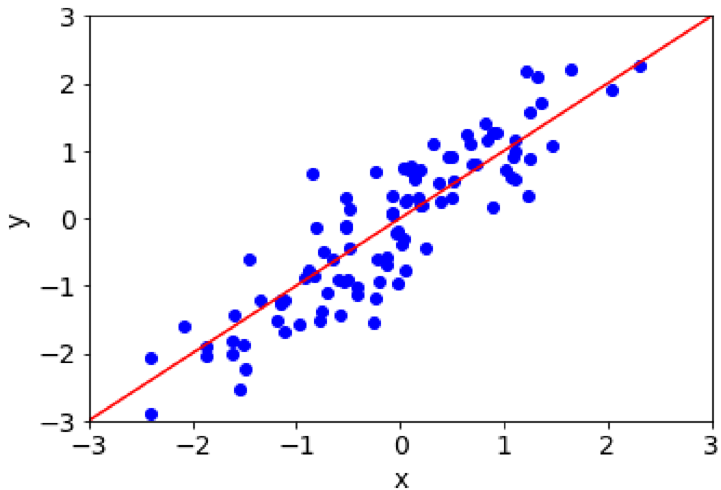
Given a **training set** $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \mathbb{R}$, find a linear function, given by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$, that minimizes the squared loss

$$L(w, b) = \sum_{i=1}^{n} (y^{(i)} - (w \cdot x^{(i)} + b))^2.$$

Is training loss a good estimate of **future** performance?

- If n is large enough: maybe.
- Otherwise: probably an underestimate.

Example



Better error estimates

Recall: k-**fold cross-validation**

- Divide the data set into k equal-sized groups $S_1, \ldots, S_k$
- For $i = 1$ to k:
  - Train a regressor on all data except S_i
  - Let E_i be its error on S_i
- Error estimate: average of $E_1, \ldots, E_k$

A nagging question:
When n is small, should we be minimizing the squared loss?

$$L(w, b) = \sum_{i=1}^{n} (y^{(i)} - (w \cdot x^{(i)} + b))^2$$

Ridge regression

Minimize squared loss **plus** a term that penalizes "complex" w:

$$L(w, b) = \sum_{i=1}^{n} (y^{(i)} - (w \cdot x^{(i)} + b))^2 + \lambda \|w\|^2$$

Adding a penalty term like this is called **regularization**.

Put predictor vectors in matrix X and responses in vector y:

$$w = (X^T X + \lambda I)^{-1} (X^T y)$$

Toy example

Training, test sets of 100 points

- $x \in \mathbb{R}^{100}$, each feature x_i is Gaussian $N(0,1)$
- $y = x_1 + \cdots + x_{10} + N(0,1)$

λ	training MSE	test MSE
0.00001	0.00	585.81
0.0001	0.00	564.28
0.001	0.00	404.08
0.01	0.01	83.48
0.1	0.03	19.26
1.0	0.07	7.02
10.0	0.35	2.84
100.0	2.40	5.79
1000.0	8.19	10.97
10000.0	10.83	12.63

The lasso

Popular "shrinkage" estimators:

- **Ridge regression**

$$L(w,b)=\sum_{i=1}^{n}(y^{(i)}-(w\cdot x^{(i)}+b))^2+\lambda\|w\|_2^2$$

- **Lasso**: tends to produce sparse w

$$L(w,b)=\sum_{i=1}^{n}(y^{(i)}-(w\cdot x^{(i)}+b))^2+\lambda\|w\|_1$$

Toy example:
Lasso recovers 10 relevant features plus a few more.

Linear models for conditional probability estimation

Topics we'll cover

1. Sources of uncertainty in prediction
2. Linear functions for conditional probability estimation
3. The logistic regression model

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.,

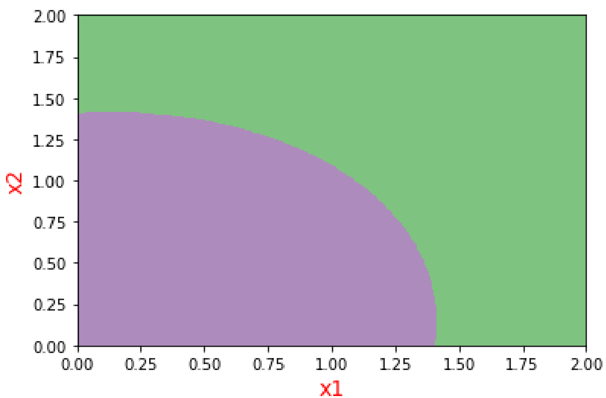
- $x =$ complete medical record for a patient at risk for a disease
- $y =$ will he/she contract the disease in the next 5 years?

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

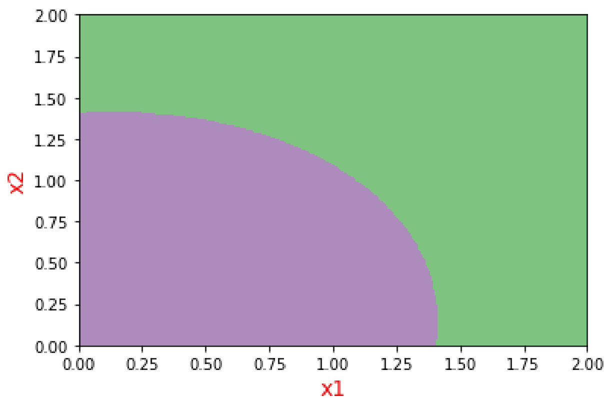
The type of classifier being used does not capture the decision boundary, e.g. using linear classifiers with:



Conditional probability estimation for binary labels

- Given: a data set of pairs (x, y), where $x \in \mathbb{R}^d$ 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 $w = (w_1, \ldots, w_d)$.

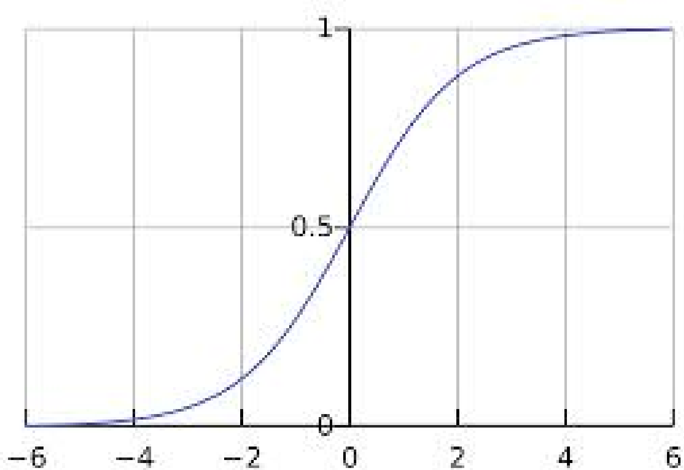
The probability of $y = 1$:

- 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

Logistic regression

Topics we'll cover

1. The logistic regression model
2. Loss function: properties
3. Solution by gradient descent

Logistic regression for binary labels

- Data $x \in \mathbb{R}^d$ and binary labels $y \in \{-1, 1\}$
- Model parametrized by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$:

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

The learning problem

Maximum-likelihood principle: given data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, 1\}$, 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**

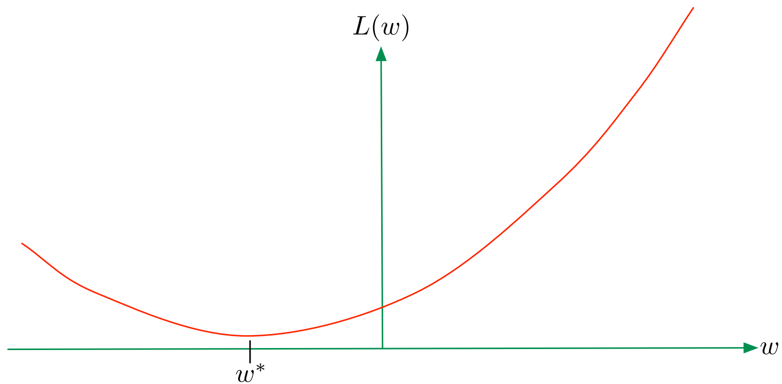
$$L(w, b) \;=\; -\sum_{i=1}^{n} \ln \Pr_{w,b}(y^{(i)} \mid x^{(i)}) \;=\; \sum_{i=1}^{n} \ln(1 + e^{-y^{(i)}(w \cdot x^{(i)} + b)})$$

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

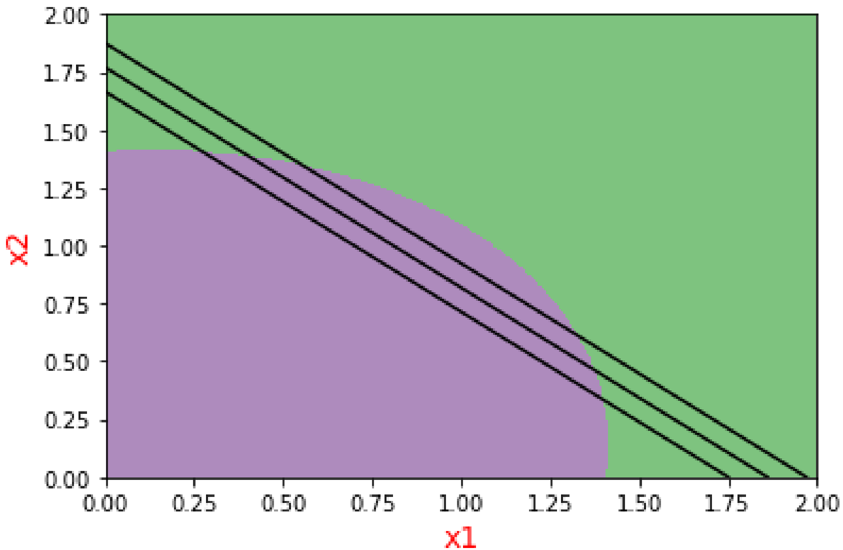
$$\arg\min_{w \in \mathbb{R}^d} L(w) \;=\; \sum_{i=1}^{n} \ln(1 + e^{-y^{(i)}(w \cdot x^{(i)})})$$

- 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



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

Logistic regression in use

Topics we'll cover

1. A text classification problem
2. Bag-of-words representation for text
3. Solution by logistic regression
4. Margin versus test error
5. Interpreting the model

Sentiment data

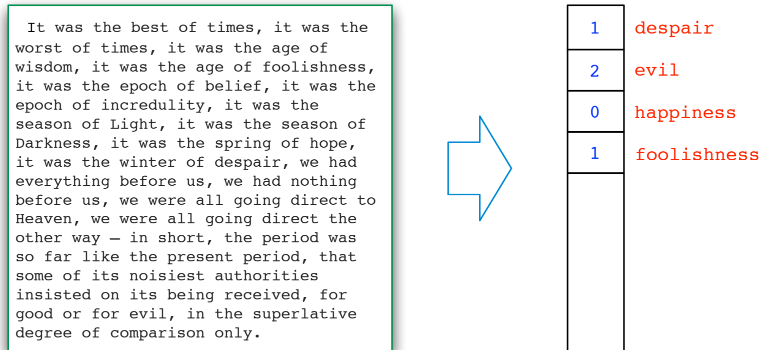
Data set: sentences from reviews on Amazon, Yelp, IMDB.
Each labeled as positive or negative.

- `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 training data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, 1\}$, find w, b minimizing

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

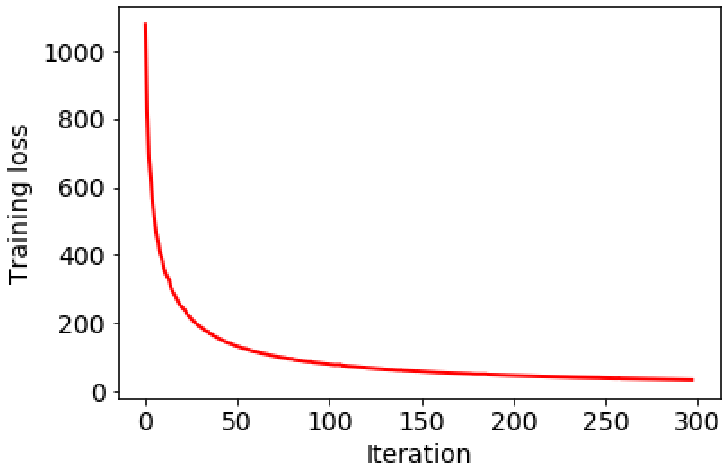
Convex problem with many solution methods, e.g.

- 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

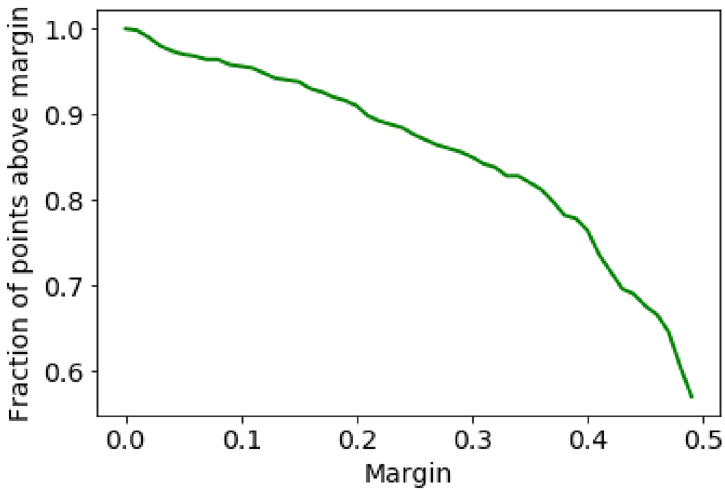
If you plan to use this in a car forget about it. -1

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

Waste your money on this game. 1

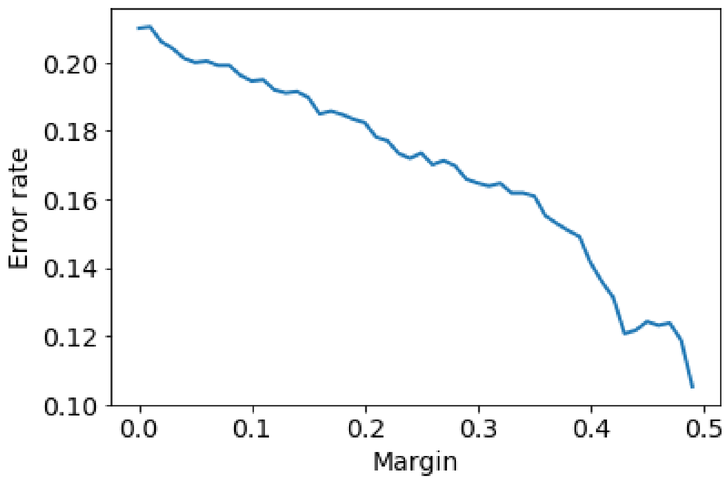
Margin and test error

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



Margin and test error

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



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'