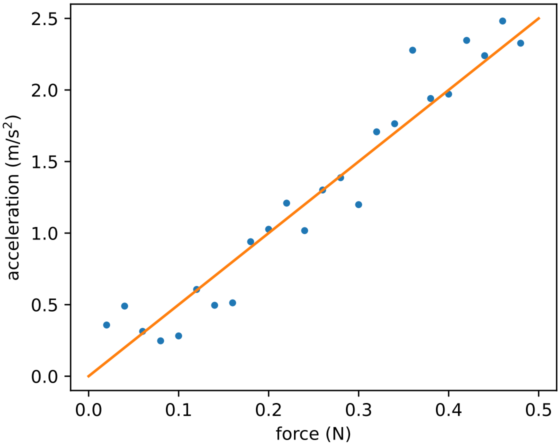
Machine Learning

Lecture 4: Classification

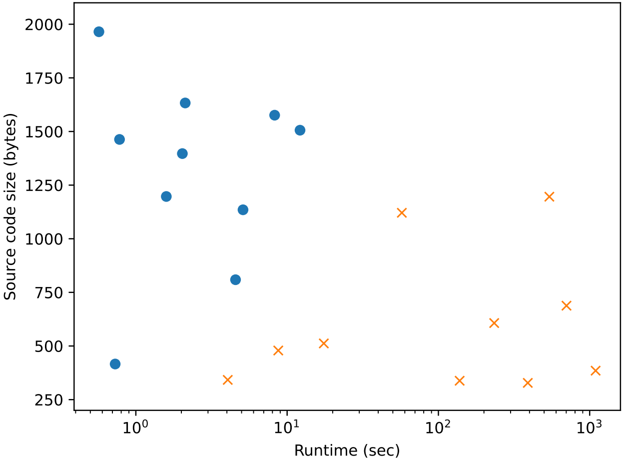
Hao Tang

October 1, 2022

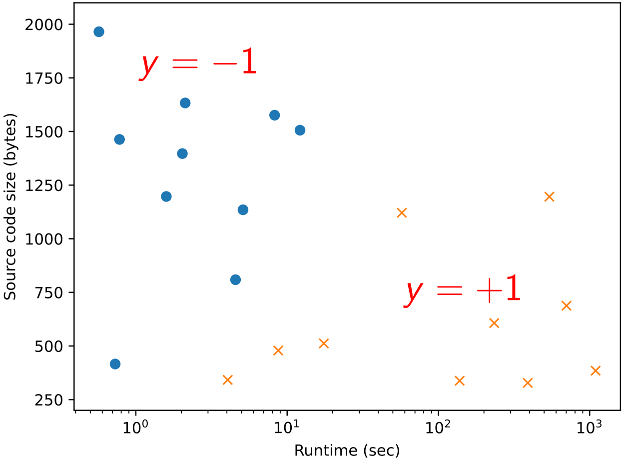
Regression



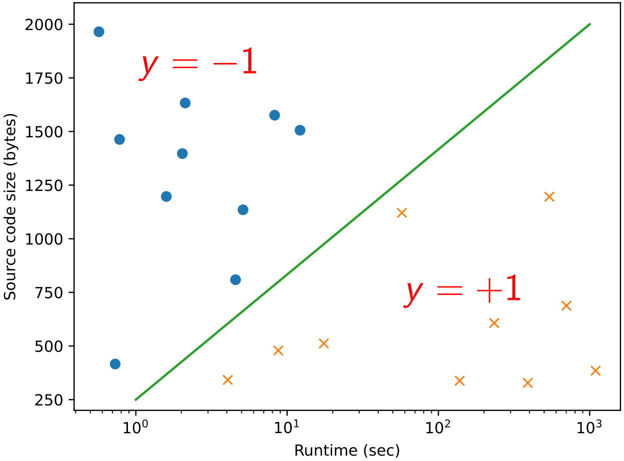
Classification



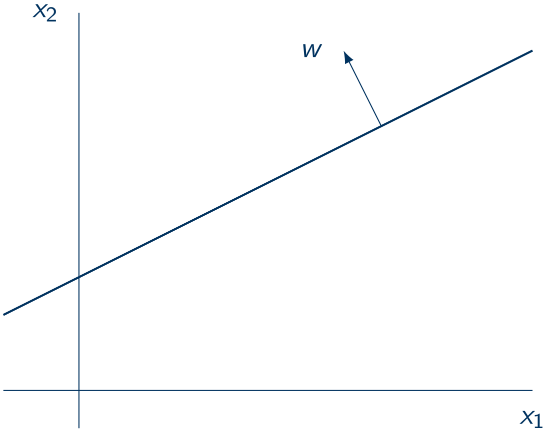
Classification



Classification

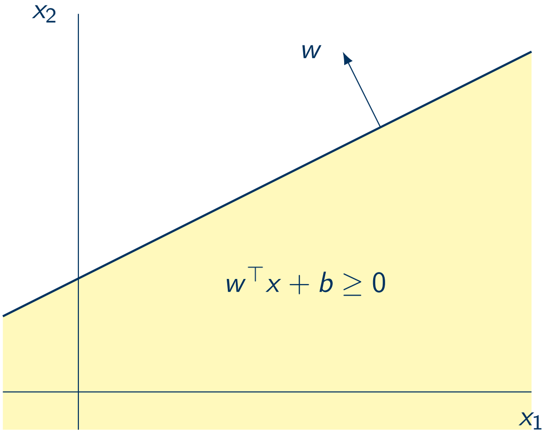


Geometry



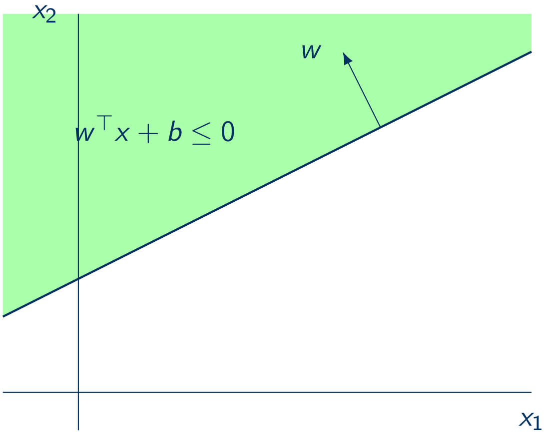
$$w^\top x + b = 0 \text{ where } x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Geometry



$$w^\top x + b = 0 \text{ where } x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Geometry



$$w^\top x + b = 0 \text{ where } x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Binary classification

$$f(x) = \begin{cases} -1 & \text{if } w^\top x + b < 0 \\ +1 & \text{if } w^\top x + b \geq 0 \end{cases} = \operatorname{sgn}(w^\top x + b) \tag{1}$$

- The plane $w^\top x + b = 0$ separates the two classes.
- The function f labels one class as 0 and the other class as 1.
- The task is called binary classification, because there are two classes.

Zero-one loss

$$\ell_{01}(\hat{y}, y) = \begin{cases} 1 & \text{if } \hat{y} \neq y \\ 0 & \text{otherwise} \end{cases} = \mathbb{1}_{\hat{y} \neq y} \tag{2}$$

- Think $\hat{y}$ as the prediction and y as the label.
- We suffer a loss of 1 if we predict the label wrong.
- In the binary case, $\ell_{01}(\hat{y}, y) = \mathbb{1}_{\hat{y}y<0}$.

Classification

- $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$: data set
  - $x = \begin{bmatrix} x[1] & \cdots & x[d] \end{bmatrix}^\top$: input, features
  - y: ground truth, label, gold reference.
- $f(x) = w^\top x + b$: linear seperator, linear predictor, hyperplane
  - $w = \begin{bmatrix} w[1] & \cdots & w[d] \end{bmatrix}^\top$: weights
  - $b \in \mathbb{R}$: bias
  - $\{w, b\}$: parameters

Classification

- Given $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, find w such that the zero-one loss

$$L = \frac{1}{N} \sum_{i=1}^{N} \ell_{01}(f(x_i), y_i) \tag{3}$$

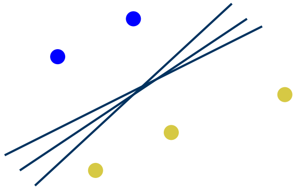
  is minimized.

- The act of finding w is called training.

- In the binary case,

$$L = \frac{1}{N} \sum_{i=1}^{N} \ell_{01}(w^\top x_i + b, y_i) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{y_i(w^\top x_i + b) < 0} \tag{4}$$

Classification



- Slightly changing w and b does not change the loss.
- The loss value only changes when the hyperplane flips the sign of a data point, and it either increases by 1 or none at all.
- The loss function (with respect to w and b) is like step functions, flat everywhere with discontinuity when the value changes.
- Finding the optimal w and b is inherently combinatorial and hard.

A probabilistic approach

- Making predictions based on signs

$$f(x) = \begin{cases} -1 & \text{if } w^\top x + b < 0 \\ +1 & \text{if } w^\top x + b \geq 0 \end{cases} = \text{sgn}(w^\top x + b) \tag{5}$$

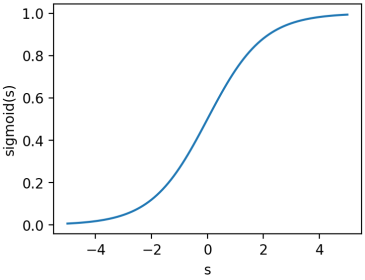
- Defining probabilities of classes

$$p(y = +1|x) = \frac{1}{1 + \exp(-(w^\top x + b))} \tag{6}$$

$$p(y = -1|x) = 1 - p(y = +1|x) \tag{7}$$

Sigmoid function

$$\sigma(s) = \frac{1}{1+\exp(-s)}$$



- When $s \to \infty$, $\sigma(s) \to 1$.
- When $s \to -\infty$, $\sigma(s) \to 0$.

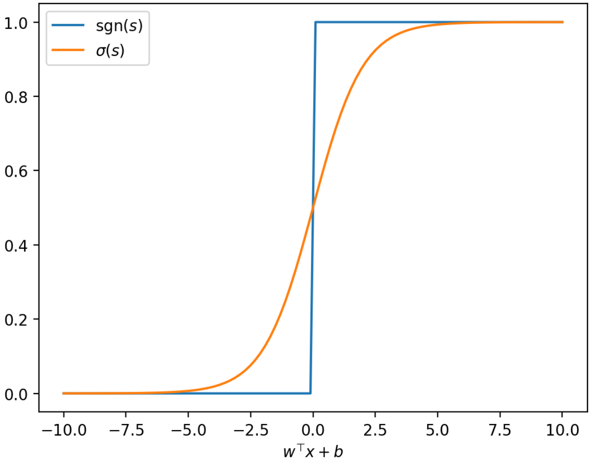
A probabilistic approach

$$p(y = +1|x) = \sigma(w^\top x + b) = \frac{1}{1 + \exp(-(w^\top x + b))} \quad (8)$$

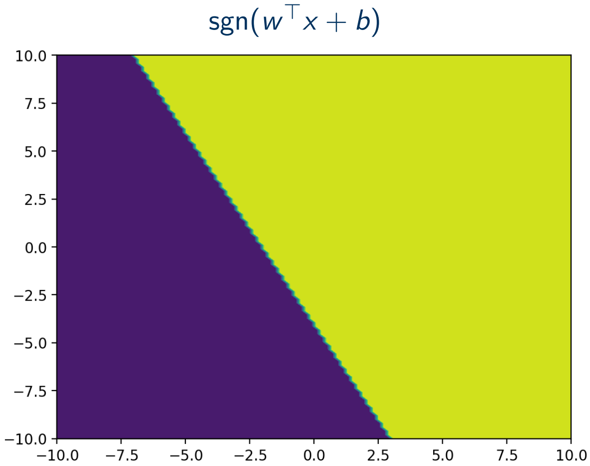
- When $w^\top x + b \to \infty$, $p(y = +1|x) \to 1$
- When $w^\top x + b \to -\infty$, $p(y = +1|x) \to 0$

$$f(x) = \begin{cases} -1 & \text{if } w^\top x + b < 0 \\ +1 & \text{if } w^\top x + b \geq 0 \end{cases} = \text{sgn}(w^\top x + b) \quad (9)$$

A probabilistic approach

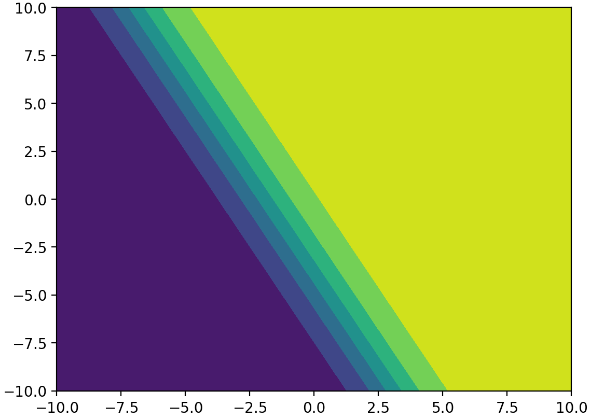


A probabilistic approach



A probabilistic approach

$$\sigma(w^\top x + b)$$



A probabilistic approach

$$p(y=+1|x) = \frac{1}{1+\exp(-(w^\top x + b))} \tag{10}$$

$$p(y=-1|x) = 1 - \frac{1}{1+\exp(-(w^\top x + b))} = \frac{\exp(-(w^\top x + b))}{1+\exp(-(w^\top x + b))} \tag{11}$$

$$= \frac{1}{\exp(w^\top x + b) + 1} \tag{12}$$

$$p(y|x) = \frac{1}{1+\exp(-y(w^\top x + b))} \tag{13}$$

Log likelihood of w and b

Given a data set $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, the likelihood of w and b is

$$L = \log \prod_{i=1}^{N} p(y_i|x_i) = \sum_{i=1}^{N} \log \frac{1}{1 + \exp(-y_i(w^\top x_i + b))} \tag{14}$$

$$= \sum_{i=1}^{N} -\log \left(1 + \exp(-y_i(w^\top x_i + b)) \right) \tag{15}$$

Log likelihood of w and b

- The zero-one loss $\sum_{i=1}^{N} \mathbb{1}_{y_i(w^\top x_i + b) < 0}$ is flat, and is hard to optimize.
- The log likelihood $L = \sum_{i=1}^{N} -\log(1 + \exp(-y_i(w^\top x_i + b))$ has curvature.
- However, unlike linear regression,

$$\frac{\partial L}{\partial w} = 0 \qquad \frac{\partial L}{\partial b} = 0 \tag{16}$$

  do not have closed-form solutions.
- We will come back to this in Lecture 8.

Classification losses

- Suppose we have a labeled data point (x, y).
- Zero-one loss

$$\mathbb{1}_{y(w^\top x+b)<0} \tag{17}$$

- Log loss

$$-\log p(y|x) = \log(1 + \exp(-y(w^\top x + b)) \tag{18}$$

Notation caveat

- The log loss notation $-\log p(y|x)$ can be misleading.
- Is y the ground truth or is it a free variable?
- What it really means is $-\log p(y = y^*|x)$ given a pair (x, y^*).
- Or $-\log p(y = y_i|x_i)$ given a pair (x_i, y_i) in a data set.

Features

$$f(x) = \begin{cases} -1 & \text{if } w^\top x + b < 0 \\ +1 & \text{if } w^\top x + b \geq 0 \end{cases} = \operatorname{sgn}(w^\top x + b) \tag{19}$$

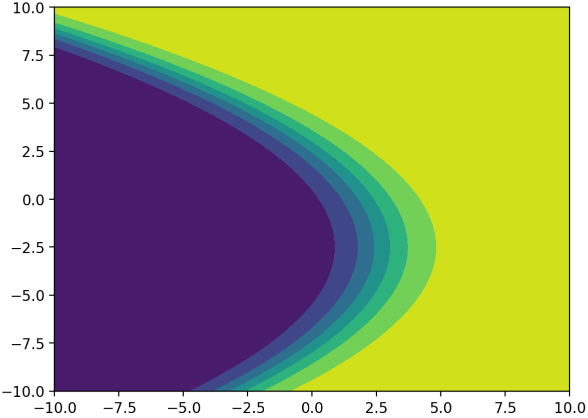
$$f(x) = \begin{cases} -1 & \text{if } w^\top \phi(x) < 0 \\ +1 & \text{if } w^\top \phi(x) \geq 0 \end{cases} = \operatorname{sgn}(w^\top \phi(x)) \tag{20}$$

Features

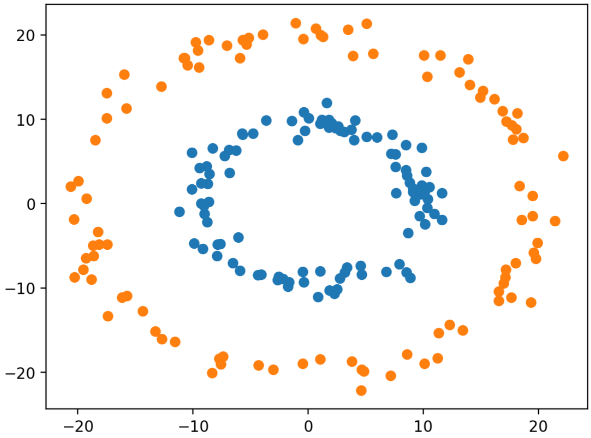
$$p(y|x) = \frac{1}{1 + \exp(-y(w^\top x + b))} \tag{21}$$

$$p(y|x) = \frac{1}{1 + \exp(-y(w^\top \phi(x)))} \tag{22}$$

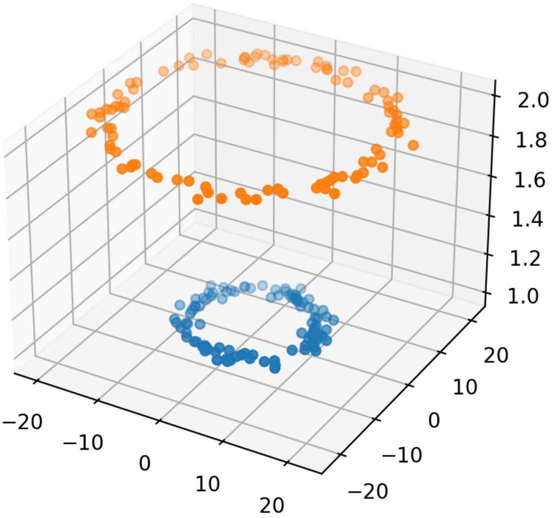
Featuers



Two-circle example



Two-circle example



Linear classification

- A linear classifier is linear in the parameters w, **not** in the features.
- A linear classifier can have arbitrary nonlinear features.

$$p(y=+1|x) = \frac{1}{1+\exp(-w^\top \phi(x))} \tag{23}$$

$$= \frac{1}{1+\exp(-(w_{+1}-w_{-1})^\top \phi(x))} \tag{24}$$

$$= \frac{\exp(w_{+1}^\top \phi(x))}{\exp(w_{+1}^\top \phi(x)) + \exp(w_{-1}^\top \phi(x))} \tag{25}$$

$$p(y=-1|x) = \frac{\exp(w_{-1}^\top \phi(x))}{\exp(w_{+1}^\top \phi(x)) + \exp(w_{-1}^\top \phi(x))} \tag{26}$$

Multiclass classification

$$p(y|x) = \frac{\exp(w_y^\top \phi(x))}{\sum_{y' \in \{0,1\}} \exp(w_{y'}^\top \phi(x))} \tag{27}$$

$$p(y|x) = \frac{\exp(w_y^\top \phi(x))}{\sum_{y' \in \mathcal{Y}} \exp(w_{y'}^\top \phi(x))} \tag{28}$$

Multiclass classification

$$f(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} p(y|x) = \operatorname*{argmax}_{y \in \mathcal{Y}} w_y^\top \phi(x) \tag{29}$$

$$f(x) = \begin{cases} -1 & \text{if } w_{-1}^\top \phi(x) > w_{+1}^\top \phi(x) \\ +1 & \text{if } w_{+1}^\top \phi(x) \geq w_{-1}^\top \phi(x) \end{cases} \tag{30}$$

$$= \begin{cases} -1 & \text{if } (w_{+1} - w_{-1})^\top \phi(x) < 0 \\ +1 & \text{if } (w_{+1} - w_{-1})^\top \phi(x) \geq 0 \end{cases} \tag{31}$$

$$= \begin{cases} -1 & \text{if } w^\top \phi(x) < 0 \\ +1 & \text{if } w^\top \phi(x) \geq 0 \end{cases} \tag{32}$$

Multiclass classification

- Log loss in the binary case

$$\sum_{i=1}^{N} \log \left(1 + \exp(y_i w^\top \phi(x_i))\right) \tag{33}$$

- Log loss in the multiclass case

$$\sum_{i=1}^{N} -w_{y_i}^\top \phi(x_i) + \log \left(\sum_{y' \in \mathcal{Y}} \exp(w_{y'}^\top \phi(x_i)) \right) \tag{34}$$

Multiclass classification

binary classification	multiclass classification
$f(x) = \begin{cases} -1 & \text{if } w^\top \phi(x) < 0 \\ +1 & \text{if } w^\top \phi(x) \geq 0 \end{cases}$	$f(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} w_y^\top \phi(x)$
$p(y \mid x) = \frac{1}{1 + \exp(-y w^\top \phi(x))}$	$p(y \mid x) = \frac{\exp(w_y^\top \phi(x))}{\sum_{y' \in \mathcal{Y}} \exp(w_{y'}^\top \phi(x))}$

Softmax

$$\text{softmax}\left(\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}\right) = \begin{bmatrix} \frac{\exp(a_1)}{\sum_{i=1}^{n} \exp(a_i)} \\ \frac{\exp(a_2)}{\sum_{i=1}^{n} \exp(a_i)} \\ \vdots \\ \frac{\exp(a_n)}{\sum_{i=1}^{n} \exp(a_i)} \end{bmatrix} \tag{35}$$

Softmax

- $\text{softmax}(\begin{bmatrix}1 & 2 & 3\end{bmatrix}^\top) = \begin{bmatrix}0.09 & 0.24 & 0.67\end{bmatrix}^\top$
- $\text{softmax}(\begin{bmatrix}100 & 200 & 300\end{bmatrix}^\top) = \begin{bmatrix}10^{-87} & 10^{-44} & 1.0\end{bmatrix}^\top$
- Softmax always returns a probability distribution.
- When the dynamic range of the input is large, the result of softmax becomes "sharp."

Softmax

- Claim: $\frac{\exp(a_{\max}/\tau)}{\sum_{i=1}^{n}\exp(a_i/\tau)} \to 1$ when $\tau \to 0$.
- That means $\frac{\exp(a_j/\tau)}{\sum_{i=1}^{n}\exp(a_i/\tau)} \to 0$ when $\tau \to 0$ for any a_j that is not the max.
- We have

$$\frac{\exp(a_m/\tau)}{\sum_{i=1}^{n}\exp(a_i/\tau)} = \frac{\exp(a_m/\tau)}{\exp(a_m/\tau) + \sum_{i\neq m}\exp(a_i/\tau)} \tag{36}$$

$$= \frac{1}{1 + \sum_{i\neq m}\exp((a_i - a_m)/\tau)} \to 1 \tag{37}$$

when $\tau \to 0$ because a_m is the largest and $a_i - a_m < 0$.

Check your understanding

- What does the function of a binary classifier look like?
- What is zero-one loss?
- What is log loss?
- What does the function of a multiclass classifier look like?
- What is softmax?