Classification
Logistic Regression

(1) HW 1 due 4/21 (last updated 4/13)

Thus far, regression:

predict a continuous value given some inputs

Given $x \in \mathbb{R}^d$ predict $y = f(x)$

$y \in \mathbb{R}$ 1, 0.1, π ...

continous

Reading Your Brain, Simple Example

$f(x_1) = $ "Person"
$f(x_2) = $ "Animal"
encoding:
"person" = 1
"Animal" = 0

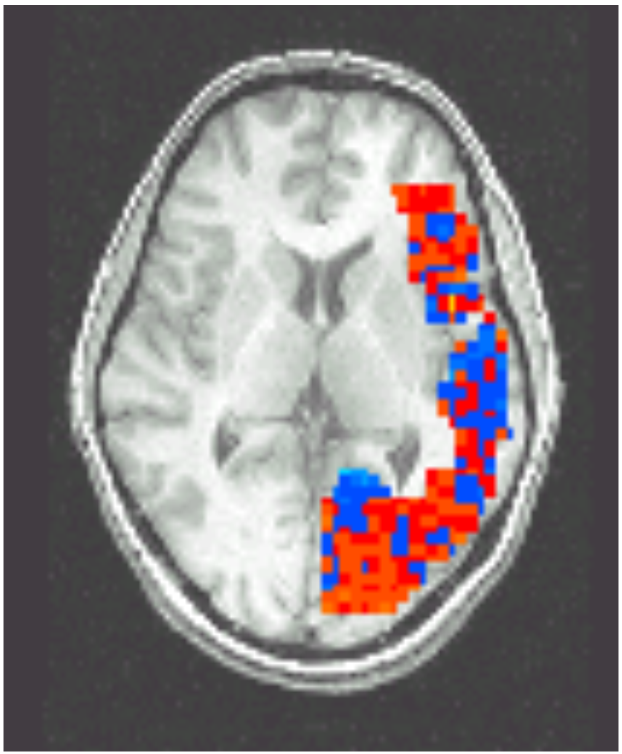
$$\begin{cases} f(x_1) = 1 \\ f(x_2) = 0 \end{cases}$$

[Mitchell et al.]

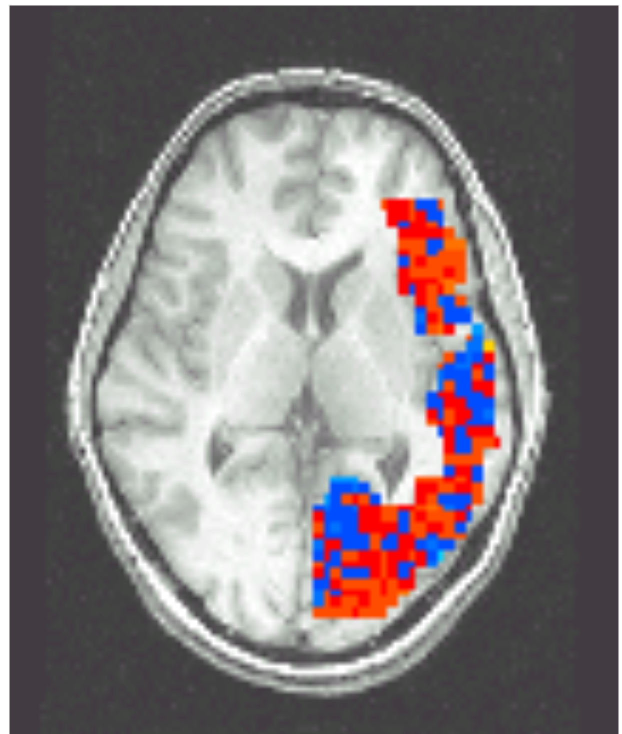
Pairwise classification accuracy: 85%

Person

Animal



x_1



x_2

Classification

- Learn f: X -> Y
  - X - features
  - Y - target classes $\underbrace{\{0,1\}}_{\text{binary}}, \underbrace{\{1,\dots,K\}}_{\text{multi-class}}$ discrete
- Loss Function $\ell(f(x), y) = 1\{f(x) \neq y\}$ $\begin{cases} 1 & \text{wrong} \\ 0 & \text{correct} \end{cases}$
- Expected loss of f:

performance measure of f

$$\mathbb{E}_{XY}[\ell(f(x), y)] = \mathbb{E}_{XY}[1\{f(x) \neq y\}] = \mathbb{E}_X \mathbb{E}_{Y|X}[1\{f(x) \neq y\} \mid X = x]$$

joint P_{XY}

$$\mathbb{E}_{Y|X}[1\{f(x) \neq Y\} \mid X = x] \overset{(\text{def})}{=} \sum_{i=1}^{K} P(Y = i \mid X = x) \cdot 1\{f(x) \neq i\}$$

$$= \sum_{i \neq f(x)} \underbrace{P(Y = i \mid X = x)}_{\text{prob of wrong pred}}$$

$$\left(\sum_{i=1}^{K} P(Y = i \mid X = x) = 1\right) \quad = 1 - \underbrace{P(Y = f(x) \mid X = x)}_{\text{Prob of correct pred}}$$

Classification

- Learn f: X -> Y
  - X - features
  - Y - target classes

- Loss Function $\ell(f(x), y) = \mathbf{1}\{f(x) \neq y\}$
- Expected loss of f:

$$\mathbb{E}_{XY}[\mathbf{1}\{f(X) \neq Y\}] = \mathbb{E}_X[\mathbb{E}_{Y|X}[\mathbf{1}\{f(x) \neq Y\}|X = x]]$$

$$\mathbb{E}_{Y|X}[\mathbf{1}\{f(x) \neq Y\}|X = x] = \sum_i P(Y = i|X = x)\mathbf{1}\{f(x) \neq i\} = \sum_{i \neq f(x)} P(Y = i|X = x)$$

$$= 1 - P(Y = f(x)|X = x)$$

- Suppose you knew P(YIX) exactly, how should you classify?

Classification

- **Learn f: X -> Y**
  - **X - features**
  - **Y - target classes**

- **Loss Function** $\ell(f(x), y) = \mathbf{1}\{f(x) \neq y\}$
- **Expected loss of f:**

$$\mathbb{E}_{XY}[\mathbf{1}\{f(X) \neq Y\}] = \mathbb{E}_X[\mathbb{E}_{Y|X}[\mathbf{1}\{f(x) \neq Y\}|X = x]]$$

$$\mathbb{E}_{Y|X}[\mathbf{1}\{f(x) \neq Y\}|X = x] = \sum_i P(Y = i|X = x)\mathbf{1}\{f(x) \neq i\} = \sum_{i \neq f(x)} P(Y = i|X = x)$$

$$= 1 - P(Y = f(x)|X = x)$$

- **Suppose you knew P(Y|X) exactly, how should you classify?**
  - **Bayes-Optimal classifier:**

based on Bayes Theorem

$$f(x) = \underbrace{\arg\max_y}_{\text{minimize all loss}} \mathbb{P}(Y = y|X = x) = \frac{P(X,Y)}{P(X)}$$

Bayes Optimal Binary Classifier

$$Y \in \{0, 1\}$$

- Suppose you knew P(Y|X) exactly, how should you classify?
  - Bayes-Optimal classifier:

$$f(x) = \arg\max_y \mathbb{P}(Y = y | X = x)$$

- Suppose we don't know P(Y|X), but have n iid examples

$$\{(x_i, y_i)\}_{i=1}^n$$

- What is a natural estimator for P(Y | X)?

Bayes Optimal Binary Classifier

- **Suppose we don't know P(Y|X), but have n iid examples**

$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ $\mathcal{D} \sim P_{XY}^n$

$Y \in \{0, 1\}$

- **What is a natural estimator for P(Y | X)?**

Fix some $\tilde{x} \in X$, (X', y) $X' \neq \tilde{X}$

Suppose $x_i = \tilde{x}$ for $m \leq n$ samples

What is a natural estimator for $\theta_* := \mathbb{P}(Y = 1|X = \tilde{x})$?

$P(Y|X) = \frac{P(X,Y)}{P(X)}$

If k of the m labels are equal to $Y = 1$ then

$k/n \approx P(X,Y)$

$\hat{P}(Y=1 | X=\tilde{X}) = \frac{k}{m}$,

$m/n \approx P(\tilde{X})$

$\mathbb{E}_{\mathcal{D}}[\hat{P}(Y=1 | X=\tilde{X})] = P(Y=1 | X=\tilde{X})$ unbiased

Bayes Optimal Binary Classifier

- **Suppose we don't know P(Y|X), but have n iid examples**

$$\{(x_i, y_i)\}_{i=1}^{n}$$

$$Y \in \{0, 1\}$$

- **What is a natural estimator for argmax_y P(Y = y | X)?**

If $X = \{0,1\}^d$, or is generally discrete

$\hat{P}(y|x)$

$$\hat{f}(x) = \arg\max_{y \in \{0,1\}} \frac{\sum_{i=1}^{n} \mathbf{1}[\mathbf{x_i = x, y_i = y}]}{\sum_{i=1}^{n} \mathbf{1}[\mathbf{x_i = x}]}$$

Issues?

(1) May not see all (x, y) pairs $\quad 2^d \cdot 2$ possibilities

(2) may not even see some x, 2^d

(3) $\hat{P}(y|x) \approx P(y|X)$ require a lot of data for each (x,y)

$\Rightarrow$ requires HUGE n

(4) continuous X

ML : use a model for $P(Y|X) \Rightarrow$ predict on unseen data

Process

Collect a **dataset** $\{(x_i, y_i)\}_{i=1}^{n}$, $y \in \{0,1\}$

Decide on a **model** $f: X \to P(y=1 \mid X)$, $f \in \mathcal{F}$

Find the function which fits the data best

Choose a loss function $\mathbb{1}\{f(x) \neq y\}$

Pick the function which minimizes loss on data $\frac{1}{n}\sum_{i=1}^{n} \ell(f(x), y) \Rightarrow f$

Use function to make prediction on new examples x_{new}

$\hat{P}(y=1 \mid X=x) = f(x_{new})$

Decide on a model, Binary Classification

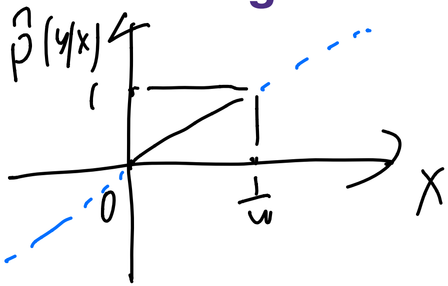
To make predictions for unseen inputs (xs),

need a **general** model for $\mathbb{P}(Y = 1|X = x)$

- **What about standard linear regression model?**

$\hat{p}(y|x) = w \cdot x$

$x \in \mathbb{R}$ $\qquad \in [0,1]$



$w \cdot x \in [-\infty, \infty]$

$\sigma(\cdot) : [-\infty, \infty] \to [0,1]$

- **Need to map real values to [0,1]**
  - **We call such maps "link functions"**

$f(x) = \sigma(w^T x)$

Logistic Regression

$z = -\infty \Rightarrow \sigma(z) = 0$

$z = \infty \Rightarrow \sigma(z) = 1$

$x \in \mathbb{R}^d, w \in \mathbb{R}^d$

Actually classification, not regression :)

Learn $\mathbb{P}(Y = 1 | X = x)$ using $\sigma(w^T x)$, for link function $\sigma =$

"S-shaped"

Logistic function(or Sigmoid): $\sigma(z) = \dfrac{1}{1 + exp(-z)}$

$g(x)$

$$\mathbb{P}[Y = 1 | X = x, w] = \sigma(w^T x) = \frac{1}{1 + \exp(-w^T x)}$$

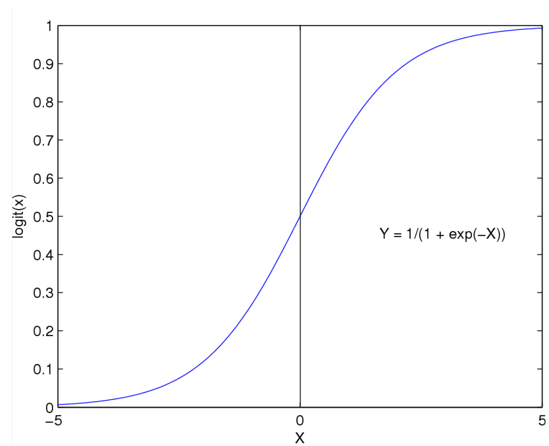
$\uparrow$

$$\mathbb{P}[Y = 0 | X = x, w] = 1 - \sigma(w^T x) = \frac{\exp(-w^T x)}{1 + \exp(-w^T x)}$$

$$= \frac{1}{1 + \exp(w^T x)}$$

$||$

1



x

Features can be discrete or continuous!

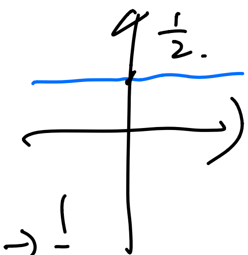
$$x \mapsto \exp(-w_j^T x), \quad j=1,\dots,K$$

$$\hat{p}(y=j \mid x) = \frac{\exp(-w_j^T x)}{\sum_{j=1}^{K} \exp(-w_j^T x)}$$

Understanding the sigmoid

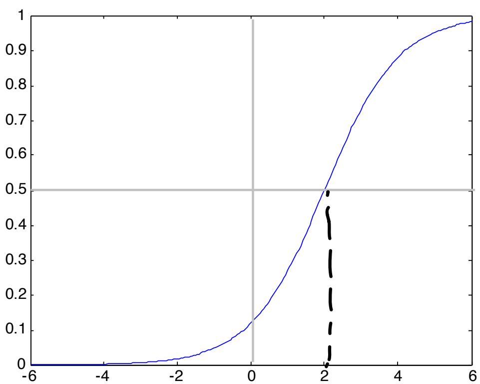
$p(y=1|x)$ $> 0.5 \Rightarrow 1$, $\leq 0.5 \Rightarrow 0$

$$\sigma(w_0 + \sum_k w_k x_k) = \frac{1}{1 + e^{-(w_0 + \sum_k w_k x_k)}}$$



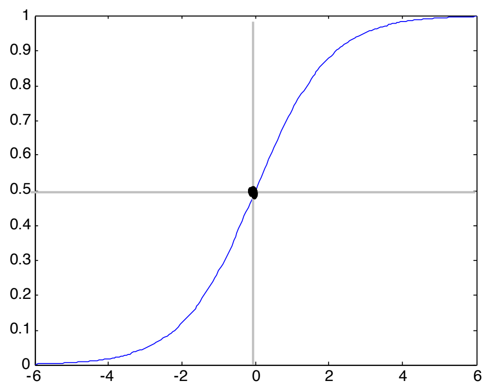
$w_0 = 0$
$w_1 \to 0 \to \frac{1}{2}$

w_0=-2, w_1=-1



$x = 2$

w_0=0, w_1=-1



x

$k=1$

w_0=0, w_1=-0.5