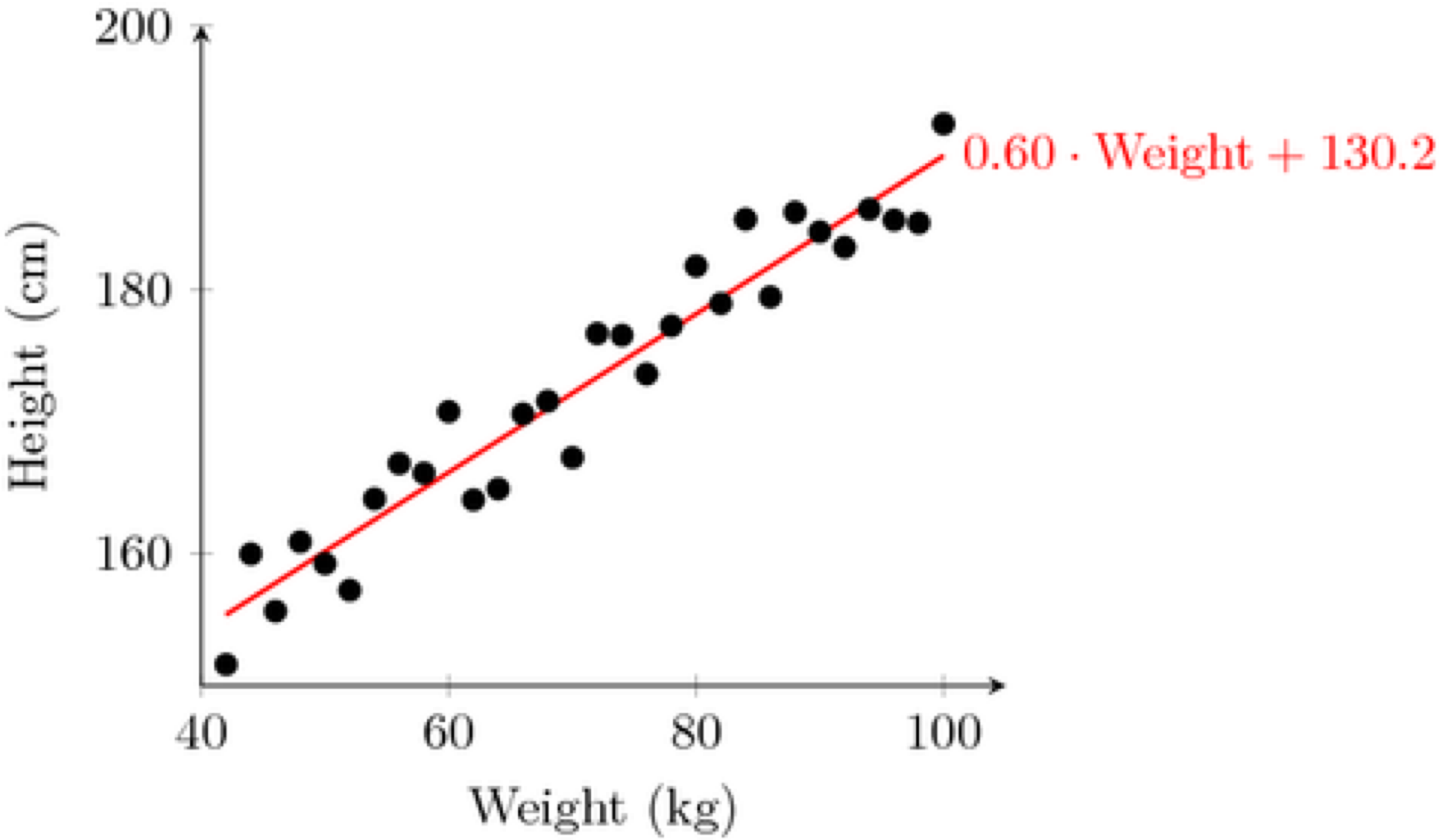
CS722/822: Machine Learning

Instructor: Jiangwen Sun
Computer Science Department

Linear Regression



Linear Regression

- Given a dataset $D = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\}$
  - $\mathbf{x}_i$ is the data vector for input variables
  - y_i is the data for the target variable
  - y_i takes numerical value
- Find a linear function f that best predicts y based on $\mathbf{x}$

$$f(\mathbf{x}_i) = \mathbf{x}_i^T \mathbf{w} \rightarrow y_i$$

- The best $\mathbf{w}$ can be found with least squares

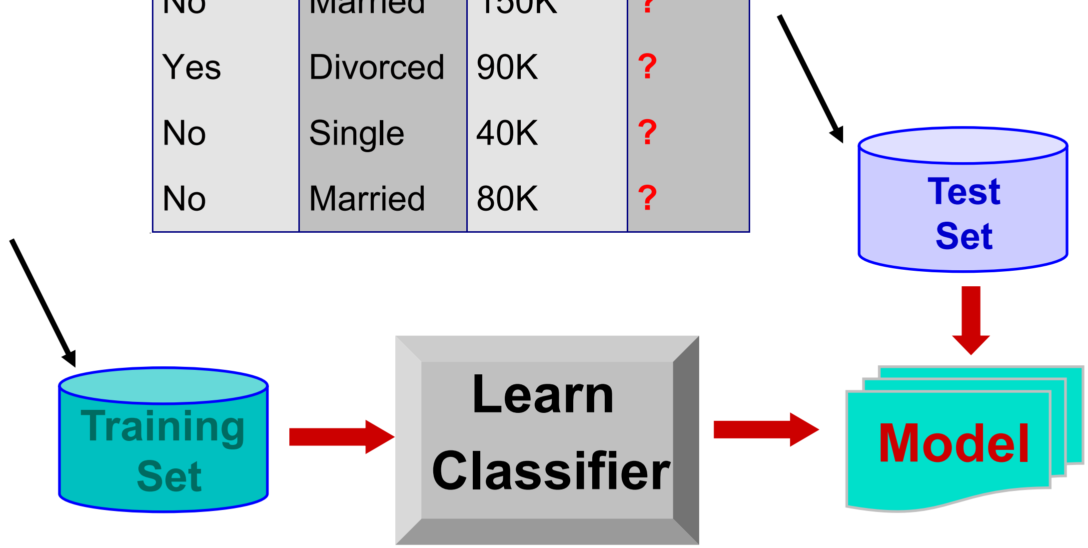
$$\min_{\mathbf{w}} \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i^T \mathbf{w}\right)^2$$

Classification (example)

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Goals: Predict if a transaction is fraud based on customer records

Why Not Linear Regression?

- Code “Yes” with 1 and “No” with -1
- Fit a linear regression model

$$y_i \leftarrow \hat{y}_i = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i$$

- Set up a threshold t (e.g., 0)
- If $\hat{y}_i \geq t$, classify $\mathbf{x}_i$ as “Yes”, or otherwise “No”

Problem

Pull all predicted values close to either 1 or -1

Restricted searching space

Suboptimal model!

Change the target

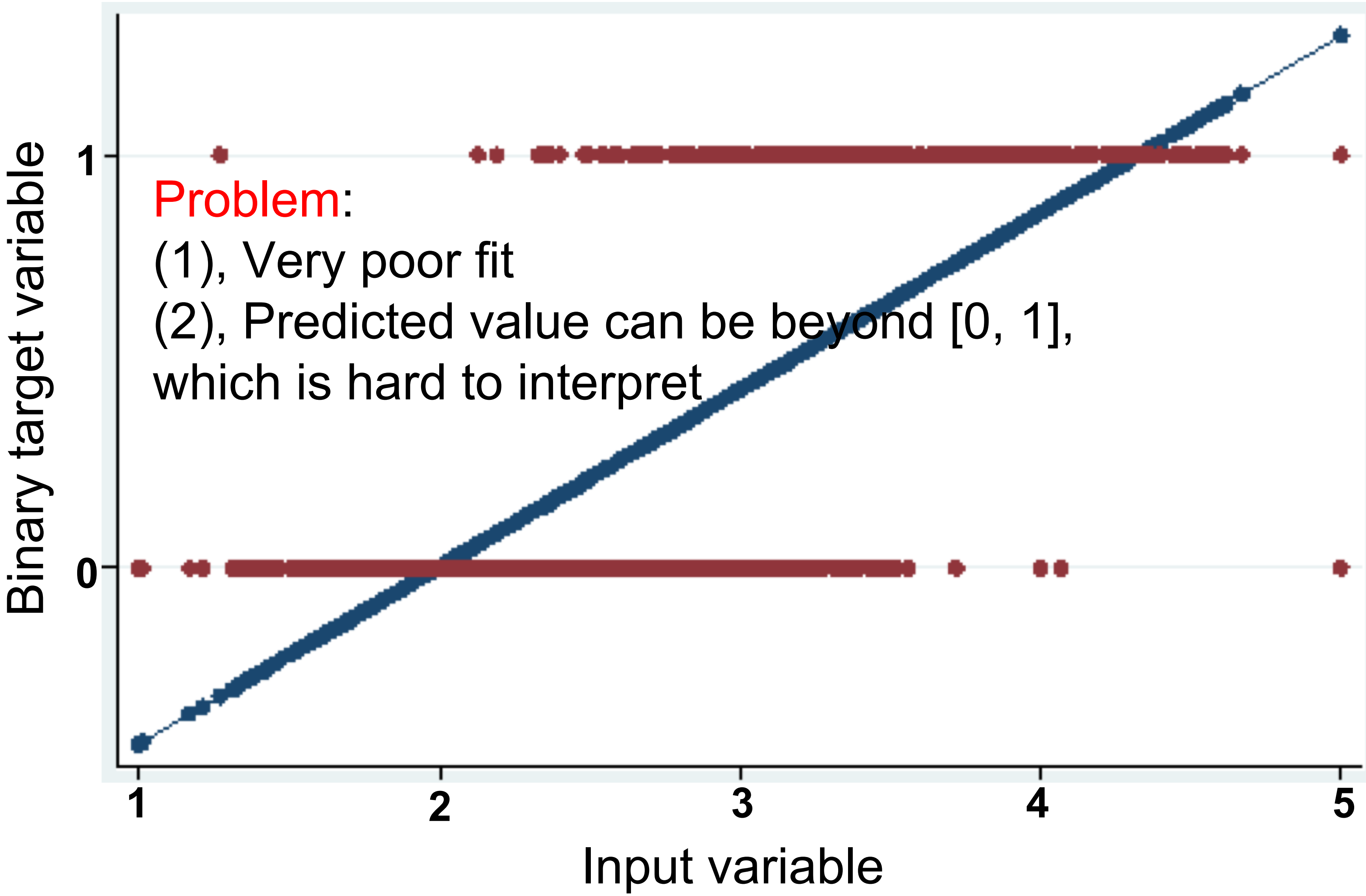
- Predict

$$p(y_i = \text{“Yes”} \,|\mathbf{x}_i)$$

- We do want the predicted value to be close to either 0 or 1

Why not linear regression, again?

Why Not Linear Regression, Again?



Logistic Regression

- Find a function f that best predicts

$$p(y_i = \text{“Yes”} \,|\mathbf{x}_i)$$

- Still find a linear function of $\mathbf{x}_i$ parameterized with $\mathbf{w}$, i.e., $\mathbf{x}_i^T\mathbf{w}$
- We want to use $\mathbf{x}_i^T\mathbf{w}$ to predict p
- $p \in [0,1]$, however, $\mathbf{x}_i^T\mathbf{w} \in (-\infty, +\infty)$

So, how to link $\mathbf{x}_i^T\mathbf{w}$ to p?

Probability to Odds

$$\text{Odds}(p) = \frac{p(y_i = \text{"Yes"}|\mathbf{x}_i)}{p(y_i = \text{"No"}|\mathbf{x}_i)} = \frac{p(y_i = \text{"Yes"}|\mathbf{x}_i)}{1 - p(y_i = \text{"Yes"}|\mathbf{x}_i)}$$

Probability $\mathrm{P}(y_i = \text{"Yes"}|\mathbf{x}_i)$	Odds
1.0	$+\infty$
0.99	99
0.75	3
0.5	1
0.25	0.3333
0.01	0.0101
0	0

$p \in [0,1]$

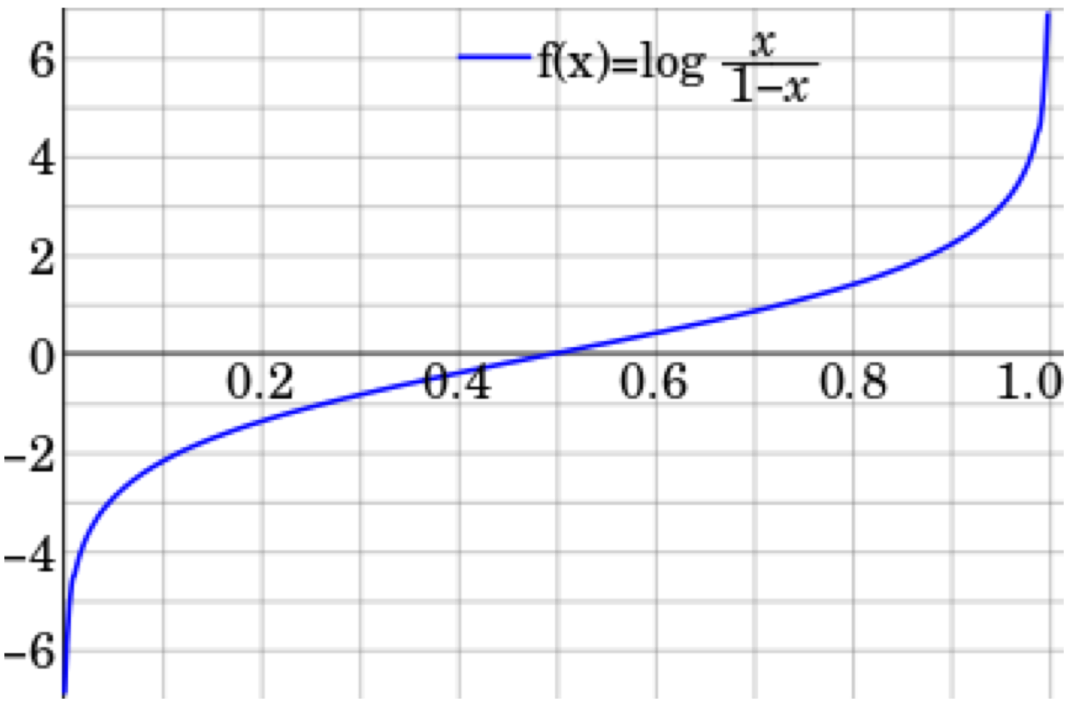
$\text{Odds}(p) \in (0, +\infty)$

$\mathbf{x}_i^T \mathbf{w} \in (-\infty, +\infty)$

Logit Function

- Take the logarithm of the odds

$$\log(Odds(p)) = \log\left(\frac{p(y_i = \text{“Yes”}|\mathbf{x}_i)}{1 - p(y_i = \text{“Yes”}|\mathbf{x}_i)}\right)$$



We call this function the **logit function** of p, i.e.,

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

$$logit(p) \in (-\infty, +\infty)$$

$$\mathbf{x}_i^T\mathbf{w} \in (-\infty, +\infty)$$

Logistic Regression

- Assume log odds is a linear function of $\mathbf{x}$

$$\log\left(\frac{p(y_i = \text{“Yes”}|\mathbf{x}_i)}{1 - p(y_i = \text{“Yes”}|\mathbf{x}_i)}\right) = \mathbf{x}_i^T \mathbf{w}$$

$$\Downarrow$$

$$p(y_i = \text{“Yes”}|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \mathbf{w})}{1 + \exp(\mathbf{x}_i^T \mathbf{w})}$$

Maximum Likelihood

- When $y_i =$ "Yes", find $\mathbf{w}$ that maximizes

$$p(y_i = \text{“Yes”} | \mathbf{x}_i)$$

- When $y_i =$ "No", find $\mathbf{w}$ that maximizes

$$p(y_i = \text{“No”} | \mathbf{x}_i) = 1 - p(y_i = \text{“Yes”} | \mathbf{x}_i)$$

- Overall

$$\prod_{i:y_i=\text{“yes”}} p(y_i = \text{“Yes”} | \mathbf{x}_i) \prod_{i:y_i=\text{“No”}} \left(1 - p(y_i = \text{“Yes”} | \mathbf{x}_i)\right)$$

Maximum Likelihood

- When $y_i =$ "Yes", find $\mathbf{w}$ that maximizes

$$p(y_i = \text{“Yes”} | \mathbf{x}_i)$$

- When $y_i =$ "No", find $\mathbf{w}$ that maximizes

$$p(y_i = \text{“No”} | \mathbf{x}_i) = 1 - p(y_i = \text{“Yes”} | \mathbf{x}_i)$$

- Overall

$$\max_{\mathbf{w}} \left(\prod_{i: y_i = \text{“yes”}} p(y_i = \text{“Yes”} | \mathbf{x}_i) \prod_{i: y_i = \text{“No”}} \left(1 - p(y_i = \text{“Yes”} | \mathbf{x}_i)\right) \right)$$

Maximum Likelihood

$$\max_{\mathbf{w}} \left(\prod_{i:y_i=\text{“yes”}} p(y_i = \text{“yes”}|\mathbf{x}_i) \prod_{i:y_i=\text{“No”}} \left(1 - p(y_i = \text{“yes”}|\mathbf{x}_i)\right) \right)$$

Let $p_i = p(y_i = \text{“yes”}|\mathbf{x}_i)$

Code “yes” with 1

Code “No” with 0

$$\max_{\mathbf{w}} \left(\prod_i p_i^{y_i} (1 - p_i)^{(1-y_i)} \right)$$

Maximum Log Likelihood

$$\max_{\mathbf{w}} \left(\prod_i {p_i}^{y_i} (1 - p_i)^{(1-y_i)} \right)$$

Take logarithm

$$\max_{\mathbf{w}} \left(\sum_i y_i \log p_i + (1 - y_i) \log(1 - p_i) \right)$$

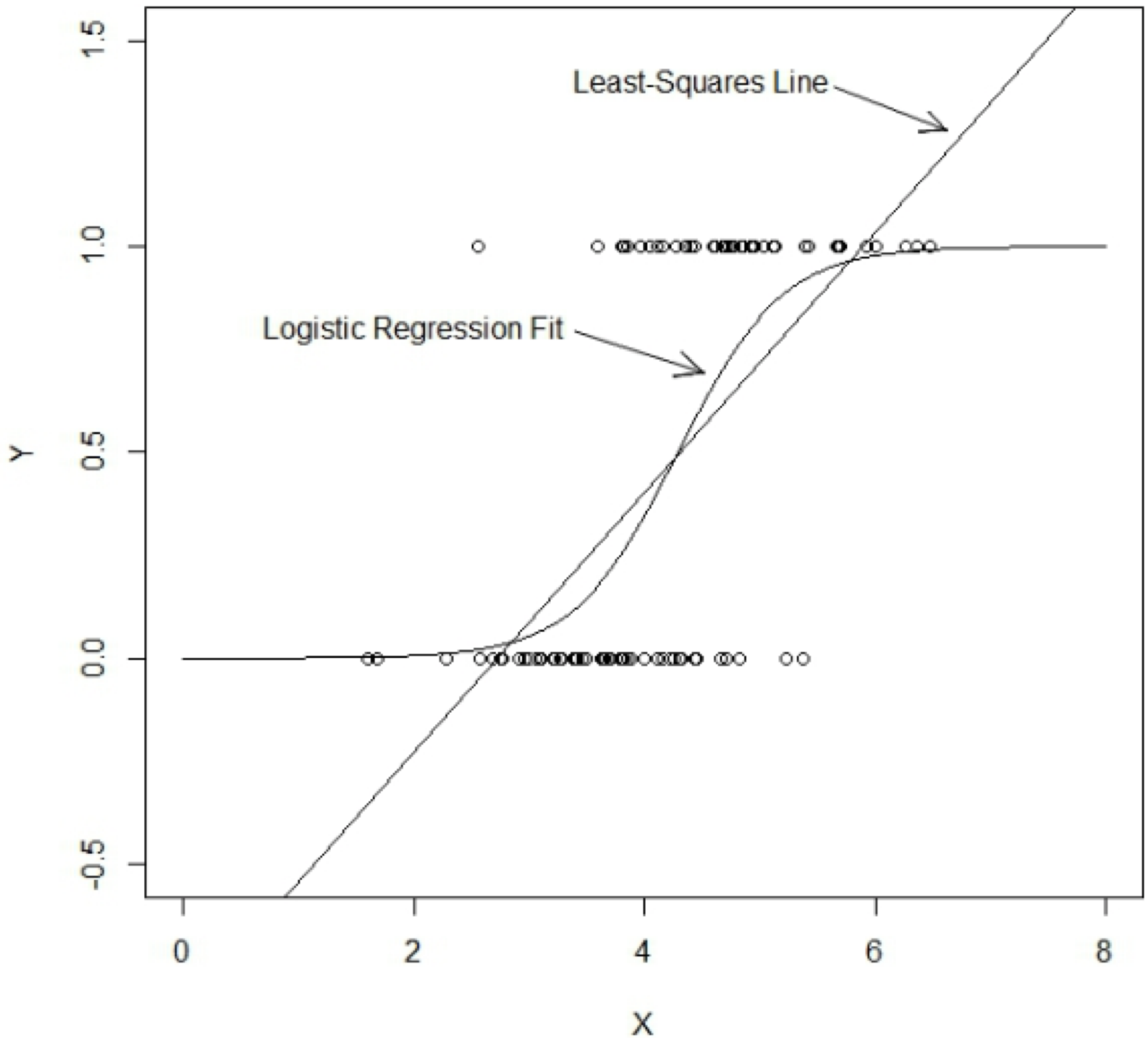
$$p_i = \frac{\exp(\mathbf{x}_i^T \mathbf{w})}{1 + \exp(\mathbf{x}_i^T \mathbf{w})}$$

$$\max_{\mathbf{w}} \left(\sum_i y_i \mathbf{x}_i^T \mathbf{w} - \log\left(1 + \exp(\mathbf{x}_i^T \mathbf{w})\right) \right)$$

$$\min_{\mathbf{w}} \left(\sum_i \log\left(1 + \exp(\mathbf{x}_i^T \mathbf{w})\right) - y_i \mathbf{x}_i^T \mathbf{w} \right)$$

Solved with gradient descent

Linear Regression VS Logistic Regression



Penalized Logistic Regression

- Ridge (Gaussian prior on $\mathbf{w}$)

$$\min_{\mathbf{w}} \left(\sum_i \left(\log\left(1 + \exp(\mathbf{x}_i^T \mathbf{w})\right) - y_i \mathbf{x}_i^T \mathbf{w} \right) + \lambda ||\mathbf{w}||^2 \right)$$

- Lasso (Laplace prior on $\mathbf{w}$)

$$\min_{\mathbf{w}} \left(\sum_i \left(\log\left(1 + \exp(\mathbf{x}_i^T \mathbf{w})\right) - y_i \mathbf{x}_i^T \mathbf{w} \right) + \lambda ||\mathbf{w}||_1 \right)$$

Multinomial Logistic Regression

Multiple Classes, say from 1 to K

- One VS all other, build K binary classifiers

$$\log\left(\frac{p(y_i = k|\mathbf{x}_i)}{1 - p(y_i = k|\mathbf{x}_i)}\right) = \mathbf{x}_i^T\mathbf{w}_k, k = 1, \dots, \mathrm{K}$$

- One VS Pivot (say, K), build K -1 binary classifiers

$$\log\left(\frac{p(y_i = k|\mathbf{x}_i)}{p(y_i = \mathrm{K}|\mathbf{x}_i)}\right) = \mathbf{x}_i^T\mathbf{w}_k, k = 1, \dots, \mathrm{K} - 1$$

Label $\mathbf{x}_i$ with class k that has the highest probability during prediction