10-701: Introduction to Machine Learning Lecture 7 – Logistic Regression

Henry Chai

2/7/24

Front Matter

- Announcements:
  - HW2 released 2/7 (today!), due on 2/16 at 11:59 PM
- Recommended Readings:
  - Murphy, Section 8.1 - 8.3

Recall: Probabilistic Learning

- Previously:
  - (Unknown) Target function, $c^*: \mathcal{X} \rightarrow \mathcal{Y}$
  - Classifier, $h : \mathcal{X} \rightarrow \mathcal{Y}$
  - Goal: find a classifier, h, that best approximates c^*
- Now:
  - (Unknown) Target *distribution*, $y \sim P^*(Y|\boldsymbol{x})$
  - Distribution, $P(Y|\boldsymbol{x})$
  - Goal: find a distribution, P, that best approximates P^*

$\vec{\theta} = [\theta_{1,0}, \theta_{1,1}, \theta_{2,0}, \theta_{2,1} \ldots]$

Recipe for Naïve Bayes

- Define a model and model parameters
  - Make the Naïve Bayes assumption
  - Assume independent, identically distributed (iid) data
  - Parameters: $\pi = P(Y = 1), \theta_{d,y} = P(X_d = 1|Y = y)$
- Write down an objective function
  - Maximize the log-likelihood
- Optimize the objective w.r.t. the model parameters
  - Solve in *closed form*: take partial derivatives, set to 0 and solve

Bernoulli Naïve Bayes

- Binary label
  - $Y \sim \text{Bernoulli}(\pi)$
  - $\hat{\pi} = {}^{N_{Y=1}}/_{N}$
    - N = # of data points
    - $N_{Y=1}$ = # of data points with label 1
- Binary features
  - $X_d | Y = y \sim \text{Bernoulli}(\theta_{d,y})$
  - $\hat{\theta}_{d,y} = {}^{N_{Y=y,\, X_d=1}}/_{N_{Y=y}}$
    - $N_{Y=y}$ = # of data points with label y
    - $N_{Y=y,\, X_d=1}$ = # of data points with label y and feature $X_d = 1$

What if some Word-Label pair never appears in our training data?

Bernoulli Naïve Bayes: Making Predictions

- Given a test data point $\boldsymbol{x}' = [x_1', \dots, x_D']^T$

$$P(Y=1 \mid x') \geq P(Y=0 \mid x')$$

$$P(Y=1 \mid x') \propto P(x' \mid Y=1)\,P(Y=1) = \hat{\pi}\left(\prod_{d=1}^{D} \hat{\theta}_{d,1}^{x_d'}\left(1-\hat{\theta}_{d,1}\right)^{1-x_d'}\right)$$

$$P(Y=0 \mid x') \propto \left(1-\hat{\pi}\right)\left(\prod_{d=1}^{D} \hat{\theta}_{d,0}^{x_d'}\left(1-\hat{\theta}_{d,0}\right)^{1-x_d'}\right)$$

What if some Word-Label pair never appears in our training data?

x_1 ("hat")	x_2 ("cat")	x_3 ("dog")	x_4 ("fish")	x_5 ("mom")	x_6 ("dad")	y (Dr. Seuss)
1	1	0	0	0	0	1
0	0	1	0	0	0	0
0	0	0	1	0	0	1
0	0	0	0	1	0	0

The Cat in the Hat gets a Dog (by ???)

- If some $\hat{\theta}_{d,y} = 0$ and that word appears in our test data $\boldsymbol{x}'$, then $P(Y = y|\boldsymbol{x}') = 0$ even if all the other features in $\boldsymbol{x}'$ point to the label being y!
- The model has been overfit to the training data…
- We can address this with a prior over the parameters!

Setting the Parameters via MAP

- Binary label
  - $Y \sim \text{Bernoulli}(\pi)$
  - $\hat{\pi} = {}^{N_{Y=1}}/_{N}$
    - N = # of data points
    - $N_{Y=1}$ = # of data points with label 1
- Binary features
  - $X_d | Y = y \sim \text{Bernoulli}(\theta_{d,y})$ and $\theta_{d,y} \sim \text{Beta}(\alpha, \beta)$
  - $\hat{\theta}_{d,y} = {}^{\left(N_{Y=y,\, X_d=1} + (\alpha - 1)\right)}/_{\left(N_{Y=y} + (\alpha-1) + (\beta-1)\right)}$
    - $N_{Y=y}$ = # of data points with label y
    - $N_{Y=y,\, X_d=1}$ = # of data points with label y and feature $X_d = 1$
    - α and β are "pseudocounts" of imagined data points that help avoid zero-probability predictions.
    - Common choice: $\alpha = \beta = 2$

Recall: Building a Probabilistic Classifier

- Define a decision rule
  - Given a test data point $\boldsymbol{x}'$, predict its label $\hat{y}$ using the *posterior distribution* $P(Y = y|X = \boldsymbol{x}')$
  - Common choice: $\hat{y} = \underset{y}{\operatorname{argmax}}\, P(Y = y|X = \boldsymbol{x}')$
- Model the posterior distribution
  - Option 1 - Model $P(Y|X)$ directly as some function of X (today!)
  - Option 2 - Use Bayes' rule (Monday):

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

Modelling the Posterior

- Suppose we have binary labels $y \in \{0,1\}$ and D-dimensional inputs $\boldsymbol{x} = [1, x_1, \dots, x_D]^T \in \mathbb{R}^{D+1}$
- **Assume**

$$P(Y = 1|\boldsymbol{x}) = \text{logit}(\boldsymbol{w}^T\boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{w}^T\boldsymbol{x})}$$

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

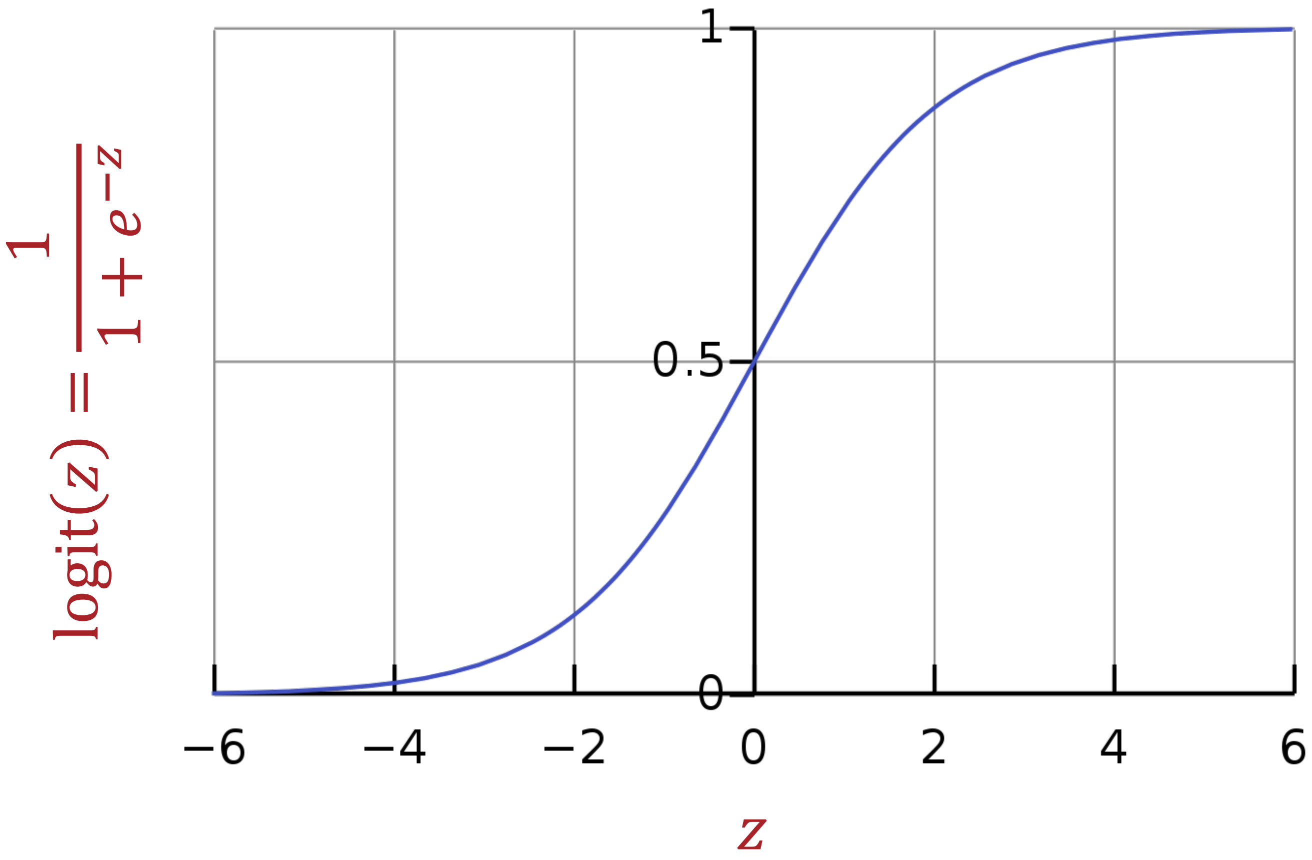
$$\frac{\exp(w^Tx)+1}{\exp(w^Tx)+1}$$

- This implies two useful facts:

1. $P(Y=0|x) = 1 - P(Y=1|x) = \frac{1}{\exp(w^Tx)+1}$

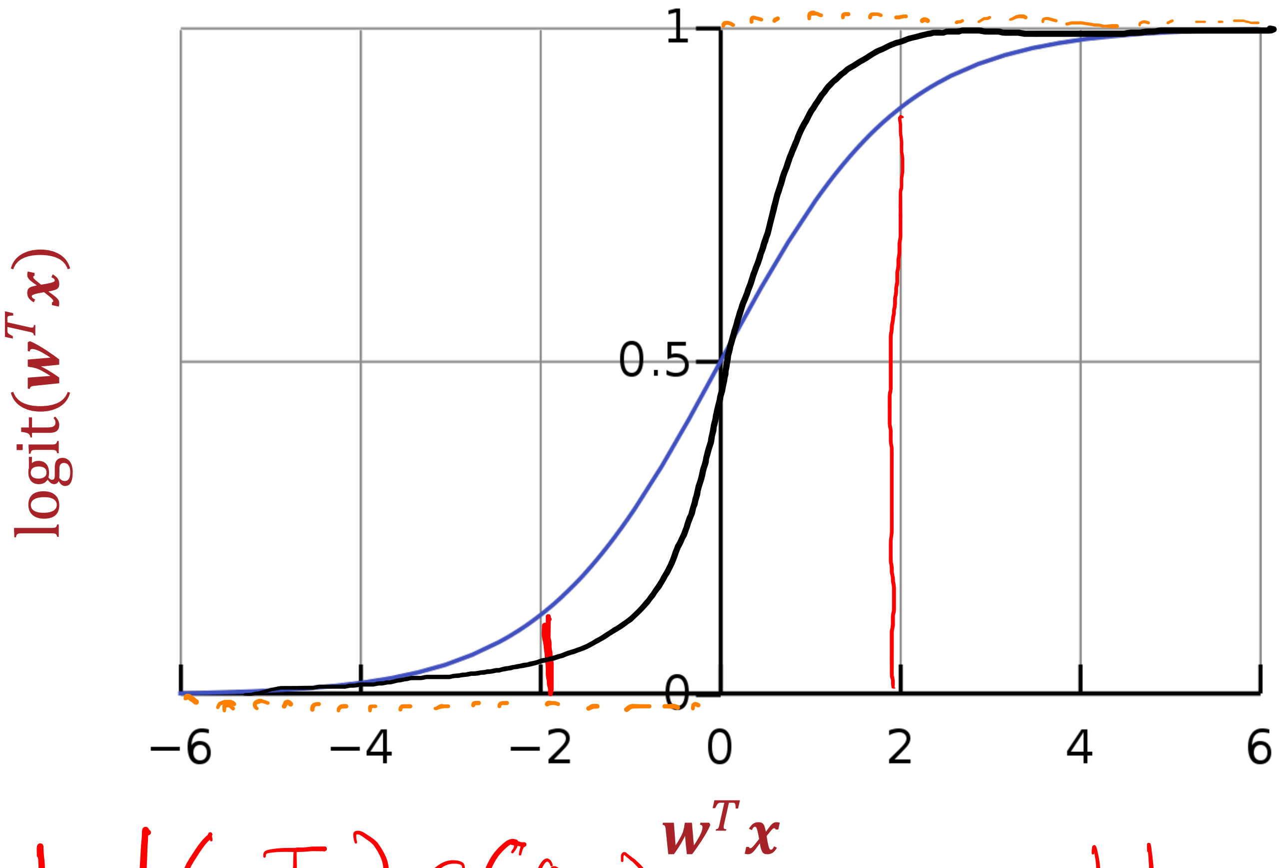
2. $\frac{P(Y=1|x)}{P(Y=0|x)} = \exp(w^Tx) \Rightarrow \log \text{odds} = w^Tx$

Logistic Function



Source: https://en.wikipedia.org/wiki/Logistic_function#/media/File:Logistic-curve.svg

Why use the Logistic Function?



logit(wTx) ∈ (0,1) ⇒ a valid probability

monotonically increasing ⇒ linear decision boundary

differentiable everywhere

Source: https://en.wikipedia.org/wiki/Logistic_function#/media/File:Logistic-curve.svg

Logistic Regression Decision Boundary

$$\hat{y} = \begin{cases} 1 \text{ if } P(Y = 1|\boldsymbol{x}') \geq \frac{1}{2} \\ 0 \text{ otherwise} \end{cases}$$

$$P(Y=1|x') = \text{logit}(\hat{w}^T x') = \frac{1}{1+\exp(-\hat{w}^T x')} \geq \frac{1}{2}$$

$$\Rightarrow 2 \geq 1 + \exp(-\hat{w}^T x')$$

$$\Rightarrow 1 \geq \exp(-\hat{w}^T x')$$

$$\Rightarrow 0 \geq -\hat{w}^T x'$$

$$\Rightarrow \hat{w}^T x' \geq 0$$

Logistic Regression Decision Boundary

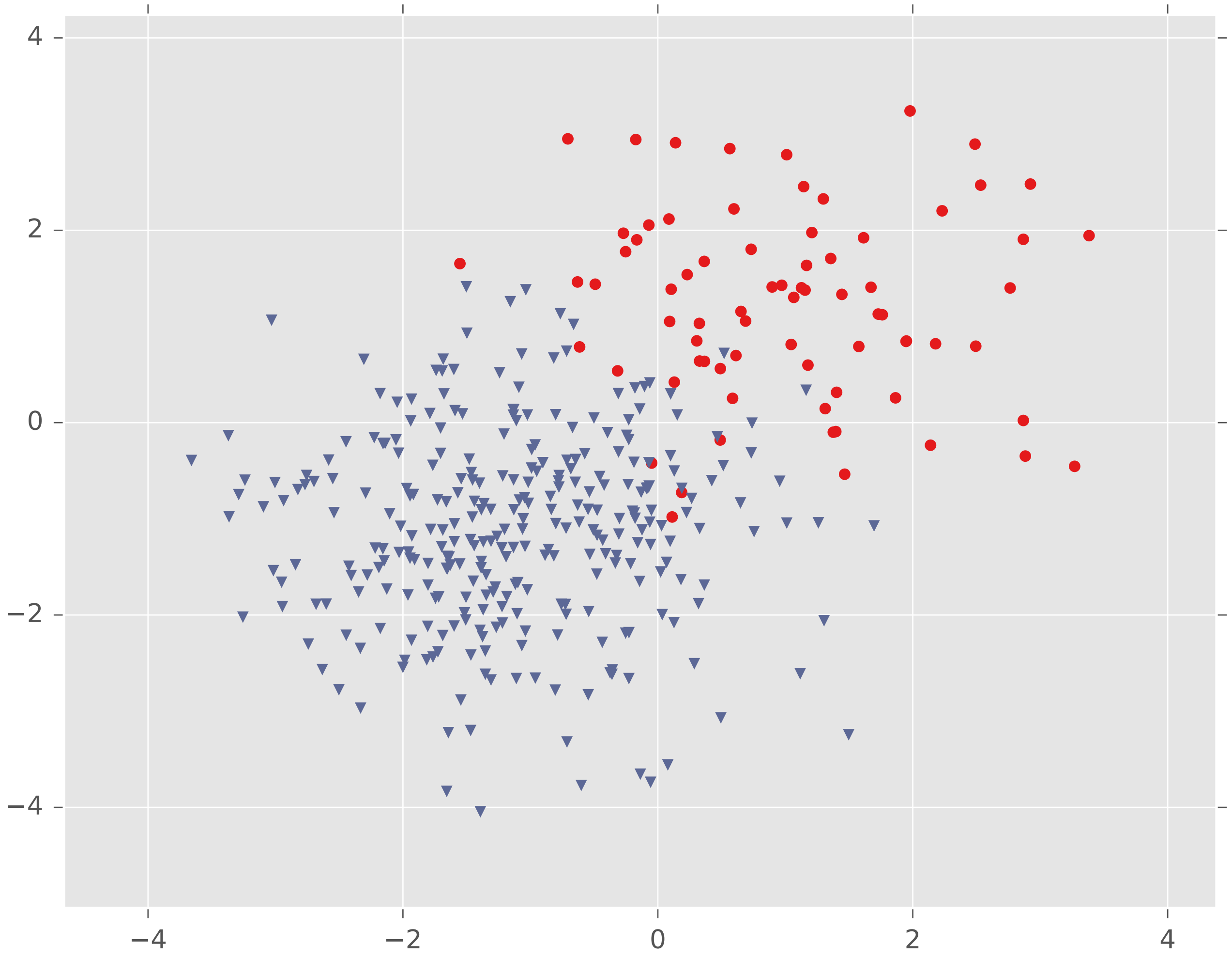


Figure courtesy of Matt Gormley

Logistic Regression Decision Boundary

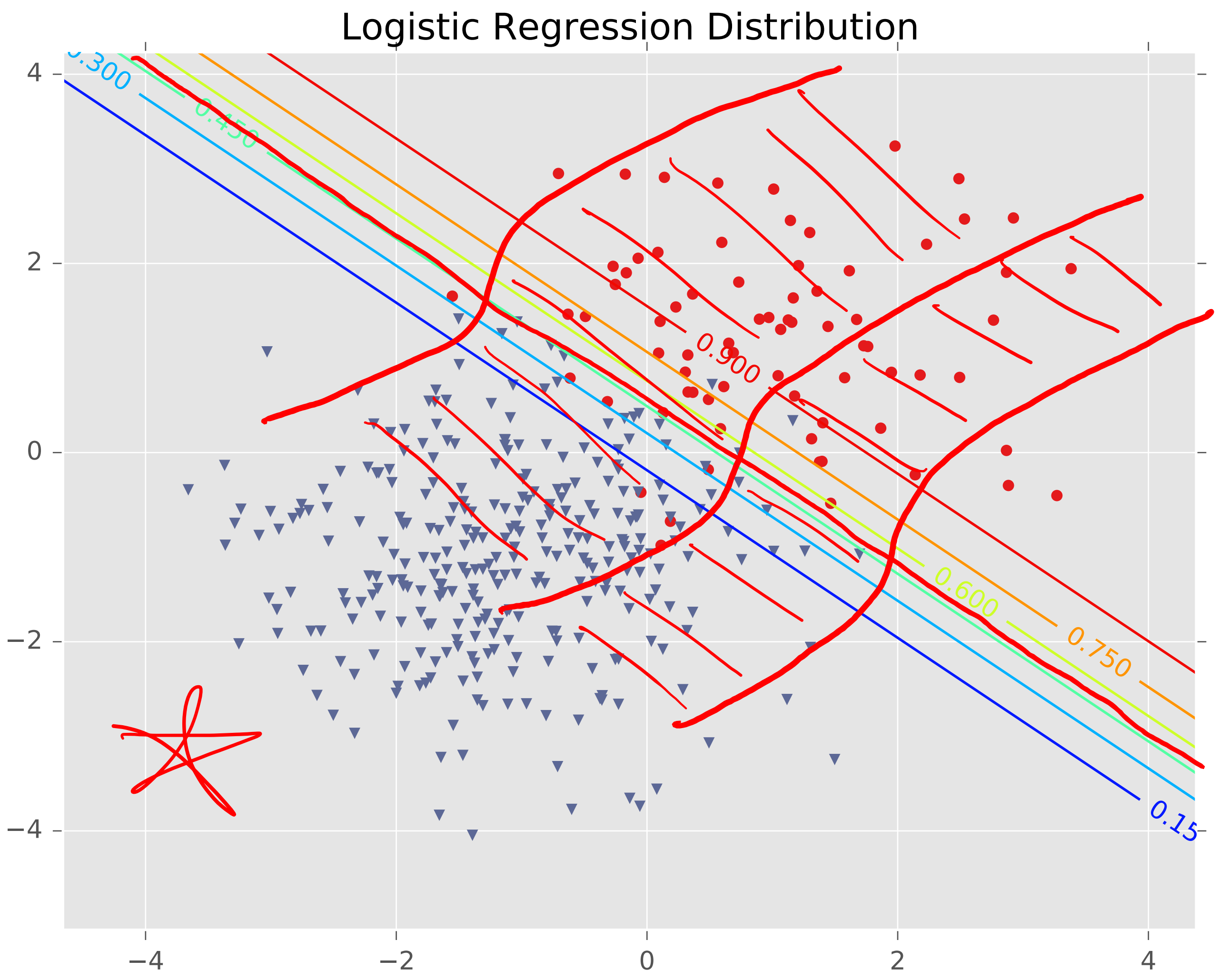


Figure courtesy of Matt Gormley

Logistic Regression Decision Boundary

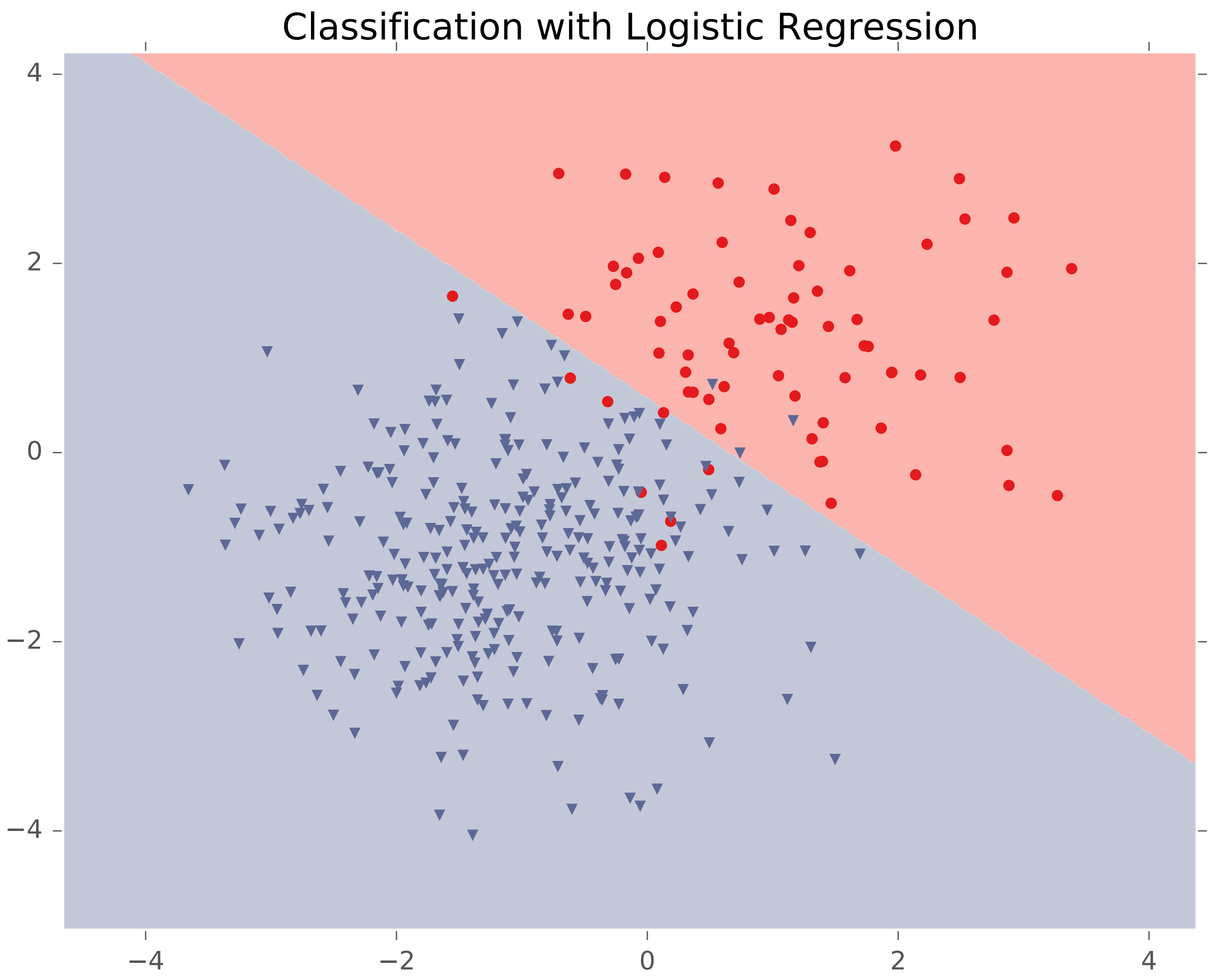


Figure courtesy of Matt Gormley

Recipe for Logistic Regression

- Define a model and model parameters
  - Assume iid data
  - Assume $P(Y=1|x) = \text{logit}(w^T x)$
  - Parameters $w = [w_0, w_1, \ldots, w_D]^T$
- Write down an objective function
  - Maximum conditional likelihood estimation
    ⇕
  - Minimum negative conditional log-likelihood estimation
- Optimize the objective w.r.t. the model parameters

???

Setting the Parameters via Minimum Negative Conditional (log-)Likelihood Estimation (MCLE)

$\log P(D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}) \ldots\})$

$= \log P(x^{(1)} \cap y^{(1)} \cap x^{(2)} \cap y^{(2)} \ldots)$

Find $\boldsymbol{w}$ that minimizes

$$\ell_{\mathcal{D}}(\boldsymbol{w}) = -\log P\left(y^{(1)}, \ldots, y^{(N)} \middle| \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}, \boldsymbol{w}\right) = -\log \prod_{n=1}^{N} P\left(y^{(n)} \middle| \boldsymbol{x}^{(n)}, \boldsymbol{w}\right)$$

$$= -\log \prod_{n=1}^{N} P(Y=1 | x^{(n)}, w)^{y^{(n)}} \left(P(Y=0 | x^{(n)}, w)\right)^{1-y^{(n)}}$$

$$= -\sum_{n=1}^{N} y^{(n)} \log\left(P(Y=1 | x^{(n)}, w)\right) + (1-y^{(n)}) \log\left(P(Y=0 | x^{(n)}, w)\right)$$

$$= -\sum_{n=1}^{N} y^{(n)} \log\left(\frac{P(Y=1 | x^{(n)}, w)}{P(Y=0 | x^{(n)}, w)}\right) + \log\left(P(Y=0 | x^{(n)}, w)\right)$$

$$= -\sum_{n=1}^{N} y^{(n)} w^T x^{(n)} + \log\left(\frac{1}{1+\exp(w^T x^{(n)})}\right)$$

$$= -\sum_{n=1}^{N} y^{(n)} w^T x^{(n)} - \log\left(1+\exp(w^T x^{(n)})\right)$$

Minimizing the Negative Conditional (log-)Likelihood

$$\ell_D(w) = -\sum_{n=1}^{N} y^{(n)} w^T x^{(n)} - \log\left(1 + \exp\left(w^T x^{(n)}\right)\right)$$

$$\nabla_w \ell_D(w) = -\sum_{n=1}^{N} y^{(n)} x^{(n)} - \frac{\exp\left(w^T x^{(n)}\right)}{1 + \exp\left(w^T x^{(n)}\right)} x^{(n)}$$

$$= \sum_{n=1}^{N} x^{(n)} \left(\frac{\exp\left(w^T x^{(n)}\right)}{1 + \exp\left(w^T x^{(n)}\right)} - y^{(n)} \right)$$

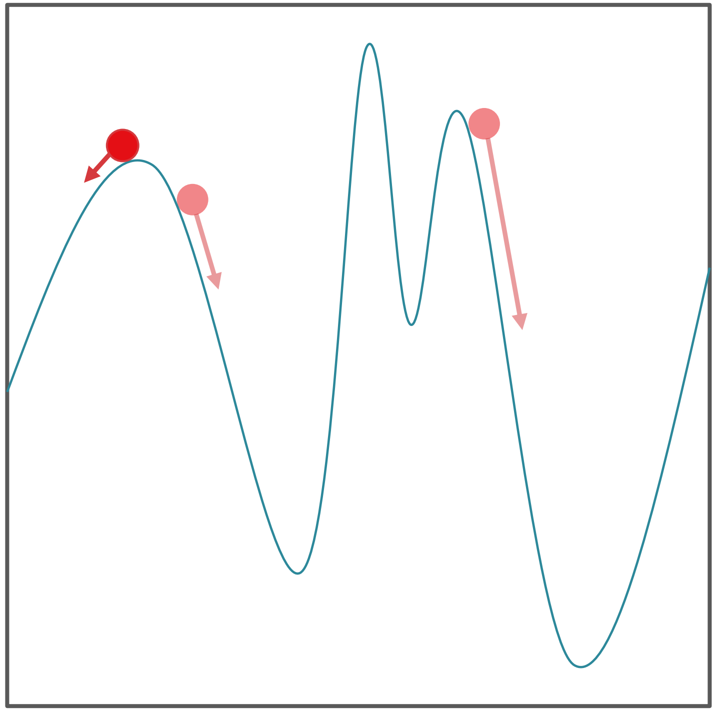
$$= \sum_{n=1}^{N} x^{(n)} \left(P\left(Y = 1 \mid x^{(n)}, w\right) - y^{(n)} \right)$$

$$w^T x^{(n)} = \sum_{d=1}^{D} w_d x_d^{(n)}$$

$$\hat{w}_d = ?$$

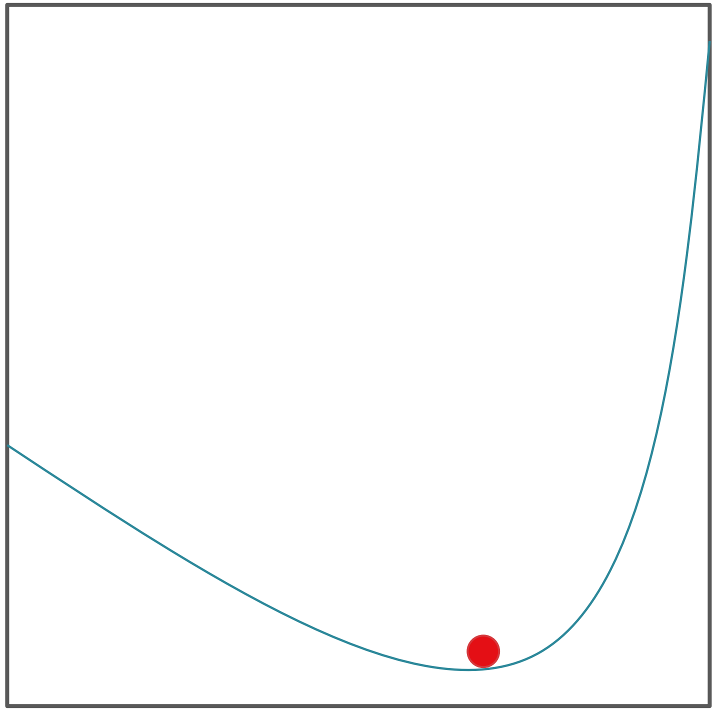
Recall: Gradient Descent

- An iterative method for minimizing functions
- Requires the gradient to exist everywhere



Recall: Gradient Descent

- An iterative method for minimizing functions
- Requires the gradient to exist everywhere



- Good news: the negative conditional log-likelihood, like the squared error, is also convex!

strictly

Gradient Descent

- Input: $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(n)}, y^{(n)}\right)\right\}_{n=1}^{N}, \eta^{(0)}$

1. Initialize $\boldsymbol{w}^{(0)}$ to all zeros and set $t = 0$
2. While TERMINATION CRITERION is not satisfied
   a. Compute the gradient:

$$O(ND)\left\{ \nabla_{\boldsymbol{w}} \ell_{\mathcal{D}}\left(\boldsymbol{w}^{(t)}\right) = \sum_{n=1}^{N} \boldsymbol{x}^{(n)}\left(P\left(Y = 1 \middle| \boldsymbol{x}^{(n)}, \boldsymbol{w}^{(t)}\right) - y^{(n)}\right)\right.$$

   b. Update $\boldsymbol{w}$: $\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)} - \eta^{(0)} \nabla_{\boldsymbol{w}} \ell_{\mathcal{D}}\left(\boldsymbol{w}^{(t)}\right)$
   c. Increment t: $t \leftarrow t + 1$

- Output: $\boldsymbol{w}^{(t)}$

Stochastic Gradient Descent

- Input: $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(n)}, y^{(n)}\right)\right\}_{n=1}^{N}, \eta_{SGD}^{(0)}$

1. Initialize $\boldsymbol{w}^{(0)}$ to all zeros and set $t = 0$
2. While TERMINATION CRITERION is not satisfied
    a. Randomly sample a data point from $\mathcal{D}$, $\left(\boldsymbol{x}^{(n)}, y^{(n)}\right)$
    b. Compute the pointwise gradient:

    $$\nabla_{\boldsymbol{w}} \ell^{(n)}\left(\boldsymbol{w}^{(t)}\right) = \boldsymbol{x}^{(n)}\left(P\left(Y = 1 \middle| \boldsymbol{x}^{(n)}, \boldsymbol{w}^{(t)}\right) - y^{(n)}\right)$$

    c. Update $\boldsymbol{w}$: $\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)} - \eta_{SGD}^{(0)} \nabla_{\boldsymbol{w}} \ell^{(n)}\left(\boldsymbol{w}^{(t)}\right)$
    d. Increment t: $t \leftarrow t + 1$

- Output: $\boldsymbol{w}^{(t)}$

Stochastic Gradient Descent

- If the data point is sampled uniformly at random, then the expected value of the pointwise gradient is proportional to the full gradient:

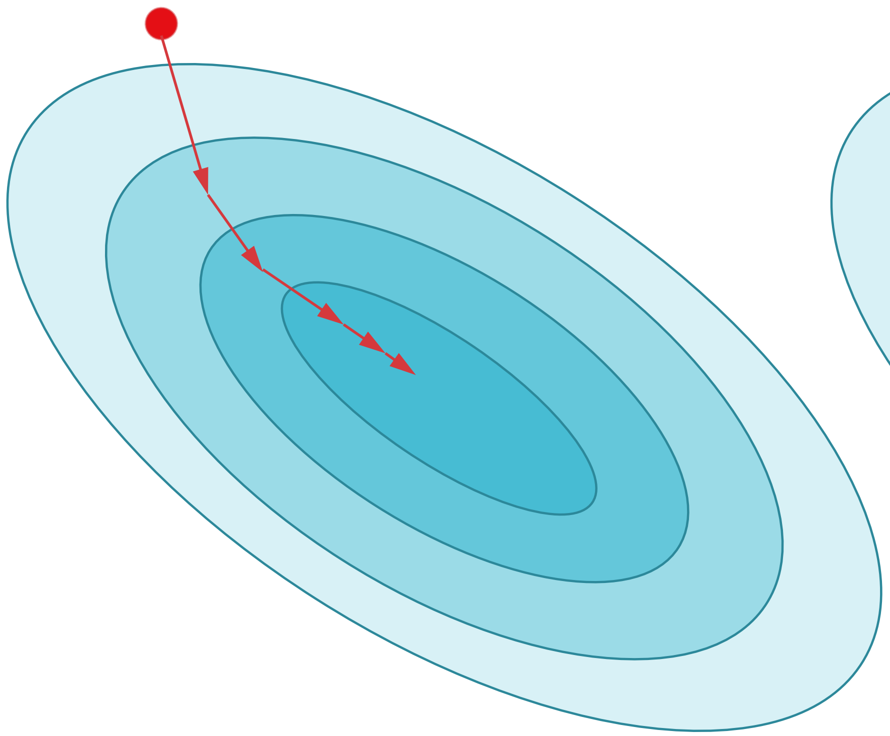
$$E\left[\nabla_{\boldsymbol{w}} \ell_{\boldsymbol{x}^{(n)}, y^{(n)}}\left(\boldsymbol{w}^{(t)}\right)\right] = \frac{1}{N}\sum_{n=1}^{N} \nabla_{\boldsymbol{w}} \ell^{(n)}\left(\boldsymbol{w}^{(t)}\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}^{(n)}\left(P\left(Y = 1 \middle| \boldsymbol{x}^{(n)}, \boldsymbol{w}^{(t)}\right) - y^{(n)}\right)$$

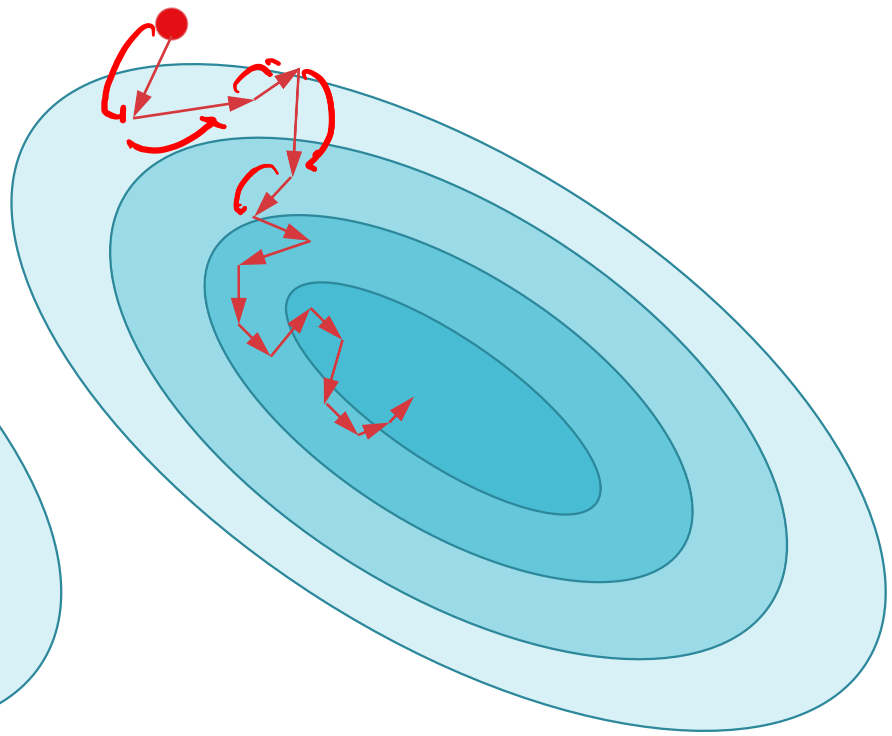
$$= \frac{1}{N} \nabla_{\boldsymbol{w}} \ell_{\mathcal{D}}\left(\boldsymbol{w}^{(t)}\right)$$

- In practice, the data set is randomly shuffled then looped through so that each data point is used equally often

Stochastic Gradient Descent vs. Gradient Descent



Gradient Descent



Stochastic Gradient Descent

Mini-batch Stochastic Gradient Descent

O(ND) → O(BD)

- Input: $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(n)}, y^{(n)}\right)\right\}_{n=1}^{N}, \eta_{MB}^{(0)}, B$

1. Initialize $\boldsymbol{w}^{(0)}$ to all zeros and set $t = 0$
2. While TERMINATION CRITERION is not satisfied
   a. Randomly sample B data points from $\mathcal{D}$:

   $$\mathcal{D}_{batch}\left\{\left(\boldsymbol{x}^{(b)}, y^{(b)}\right)\right\}_{b=1}^{B}$$

   b. Compute the gradient w.r.t. the sampled *batch*:

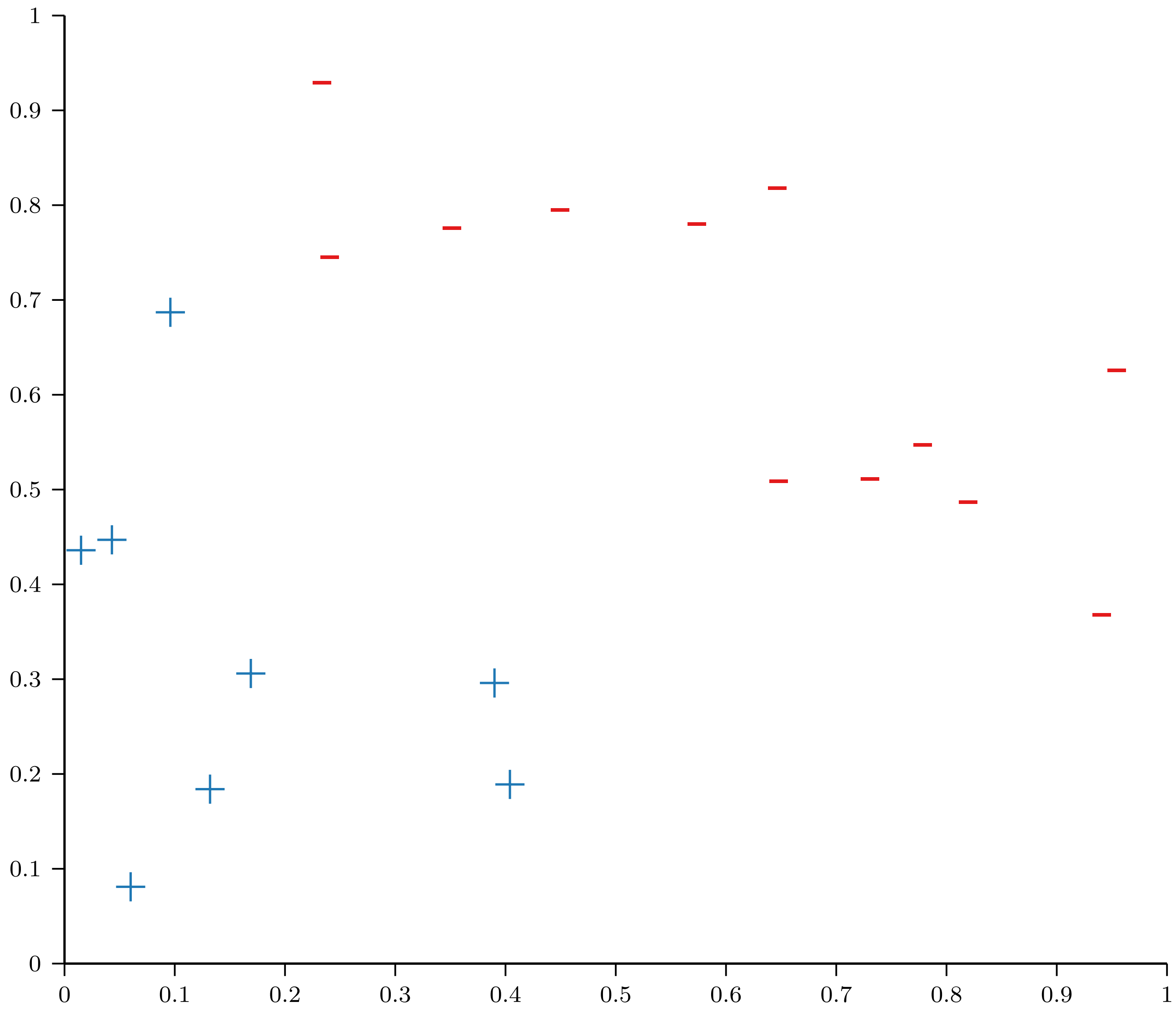
   $$\nabla_{\boldsymbol{w}} \ell_{\mathcal{D}_{batch}}\left(\boldsymbol{w}^{(t)}\right) = \sum_{b=1}^{B} \boldsymbol{x}^{(b)}\left(P\left(Y = 1 \middle| \boldsymbol{x}^{(b)}, \boldsymbol{w}\right) - y^{(b)}\right)$$

   c. Update $\boldsymbol{w}$: $\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)} - \eta_{MB}^{(0)} \nabla_{\boldsymbol{w}} \ell_{\mathcal{D}_{batch}}\left(\boldsymbol{w}^{(t)}\right)$

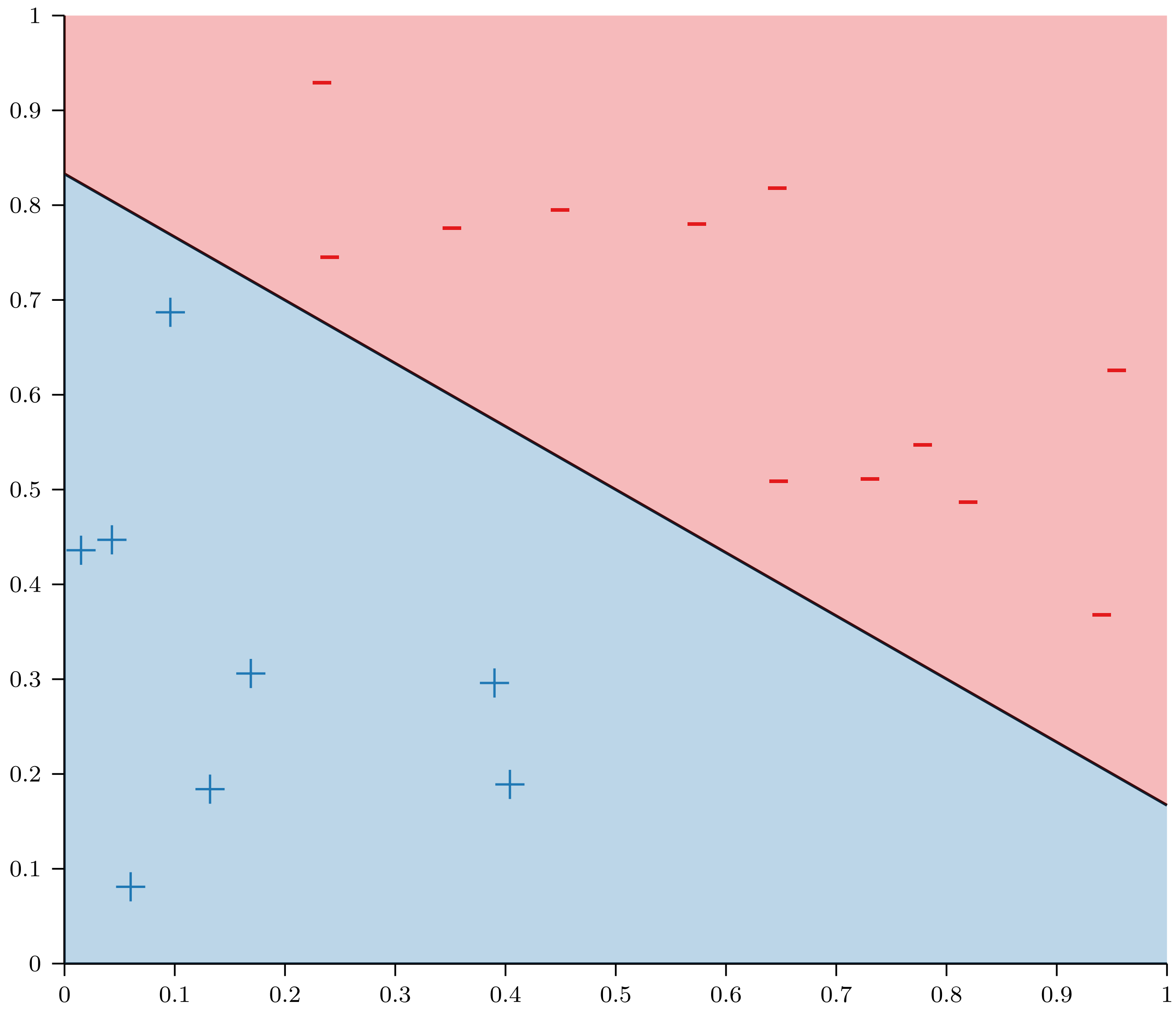
   d. Increment t: $t \leftarrow t + 1$

- Output: $\boldsymbol{w}^{(t)}$

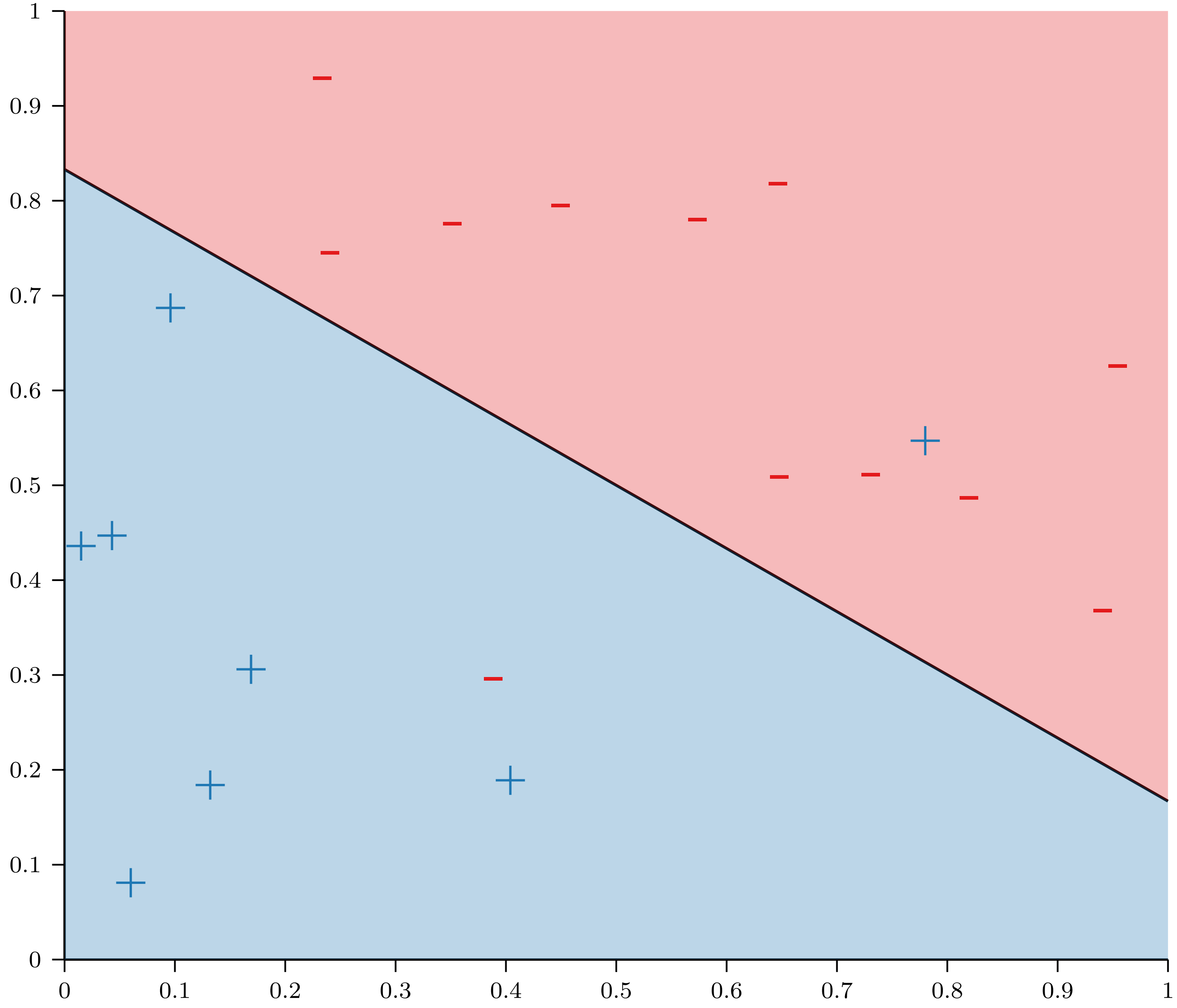
Linear Models



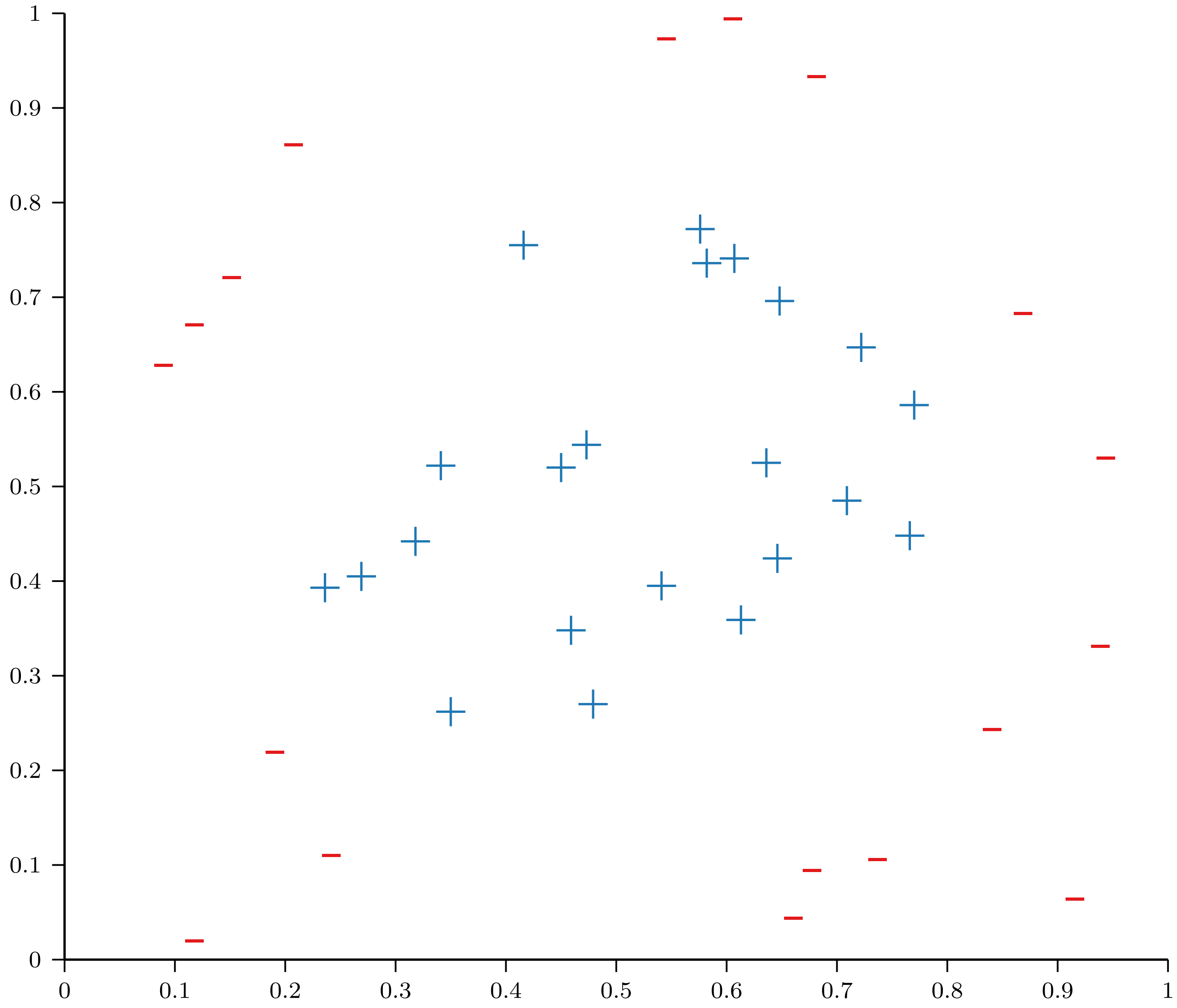
Linear Models



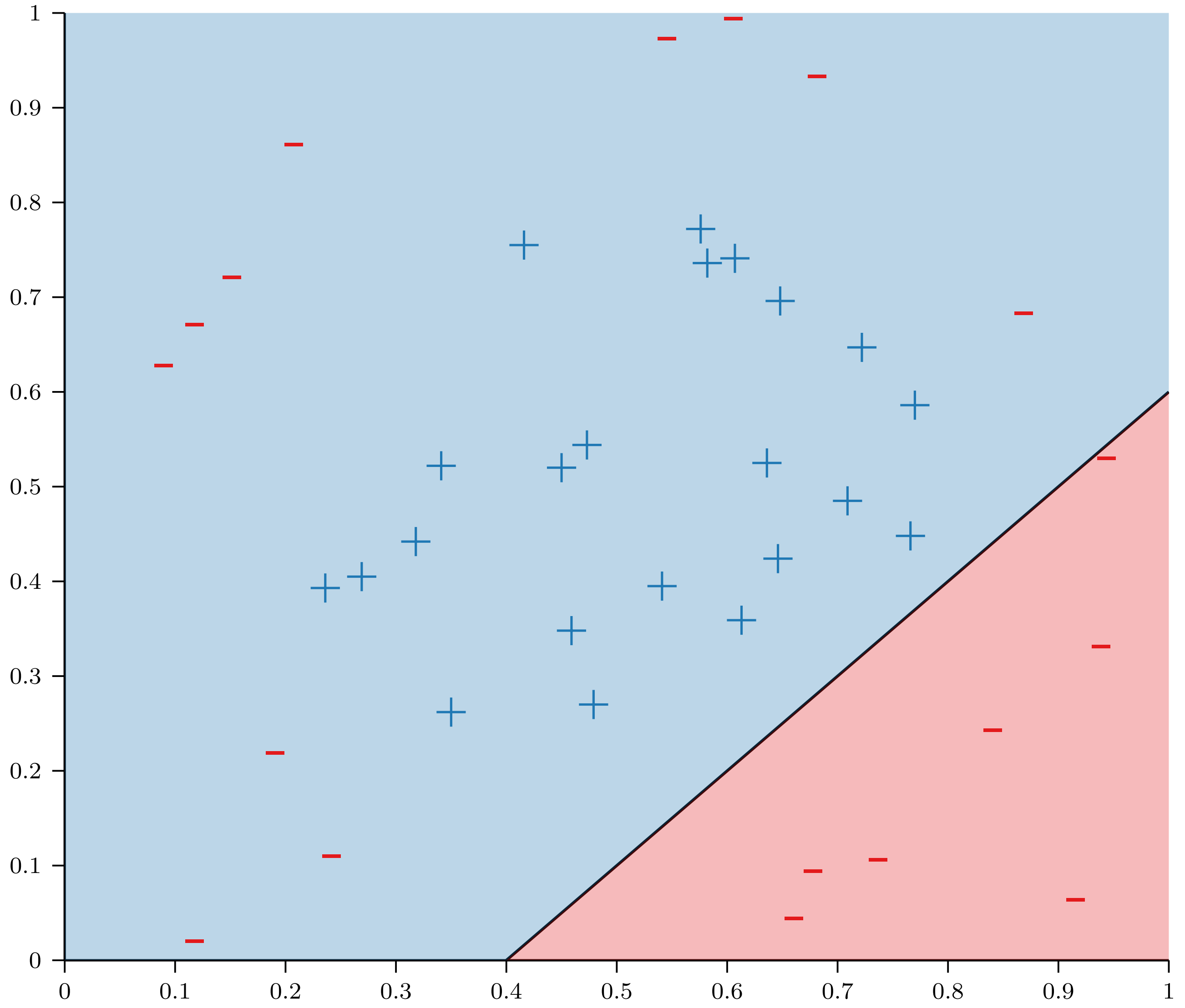
Linear Models



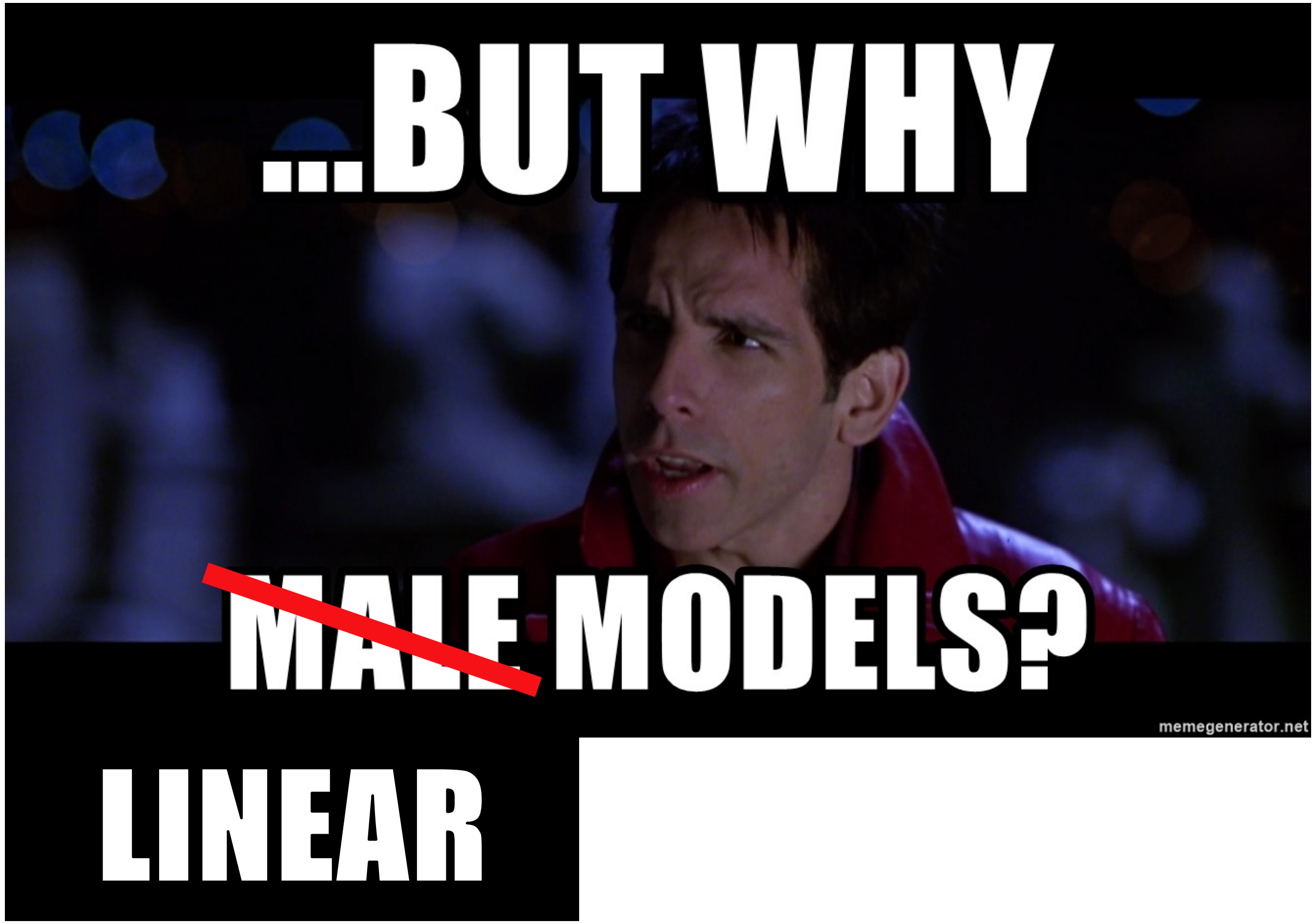
Linear Models?



Linear Models?



Linear Models?



Nonlinear Models