CS 360: Machine Learning

Prof. Sara Mathieson

Fall 2019

HAVERFORD COLLEGE

Admin

- No office hours Friday (but feel free to make an appointment)
- **Lab 5 due October 22** (Tuesday after fall break)
  - Code reviews today

Reading Quiz 5

1. The output of logistic regression is a model that creates:

   (a) a linear decision boundary

   (b) a logistic decision boundary

   (c) no decision boundary

Reading Quiz 5

1. The output of logistic regression is a model that creates:

   (a) a linear decision boundary

   (b) a logistic decision boundary

   (c) no decision boundary

2. We use logistic regression for:

   (a) classification

   (b) regression

   (c) both

Reading Quiz 5

1. The output of logistic regression is a model that creates:

   (a) a linear decision boundary

   (b) a logistic decision boundary

   (c) no decision boundary

2. We use logistic regression for:

   (a) classification

   (b) regression

   (c) both

3. Our hypothesis in logistic regression is:

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}$$

   If $\boldsymbol{w}$ is the zero vector (as it would be when starting SGD), what is the probability $y = 1$?

Reading Quiz 5

1. The output of logistic regression is a model that creates:

   (a) a linear decision boundary
   (b) a logistic decision boundary
   (c) no decision boundary

2. We use logistic regression for:

   (a) classification
   (b) regression
   (c) both

3. Our hypothesis in logistic regression is:

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}$$

   If $\boldsymbol{w}$ is the zero vector (as it would be when starting SGD), what is the probability $y = 1$?

   ½

4. How did we define the cost function for logistic regression? (Bonus: write down the cost function)

   (a) likelihood
   (b) log likelihood
   (c) negative log likelihood

Reading Quiz 5

1. The output of logistic regression is a model that creates:

   (a) a linear decision boundary
   (b) a logistic decision boundary
   (c) no decision boundary

2. We use logistic regression for:

   (a) classification
   (b) regression
   (c) both

3. Our hypothesis in logistic regression is:

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}$$

   If $\boldsymbol{w}$ is the zero vector (as it would be when starting SGD), what is the probability $y = 1$?

   ½

4. How did we define the cost function for logistic regression? (Bonus: write down the cost function)

   (a) likelihood
   (b) log likelihood
   (c) negative log likelihood

Outline for October 10

- Recap SGD for logistic regression
- Regularization
- Multi-class logistic regression
- Begin: evaluation metrics

Outline for October 10

- Recap SGD for logistic regression
- Multi-class logistic regression
- Regularization
- Begin: evaluation metrics

Stochastic Gradient Descent for Logistic Regression (binary classification)

```
set w = 0 vector
while cost J(w) still changing:
    shuffle data points
    for i = 1…n:
        w <- w – alpha(derivative of J(w) wrt x_i)
    store J(w)
```

3 important pieces to SGD

- Hypothesis function (prediction)

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}$$

3 important pieces to SGD

- Hypothesis function (prediction)

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}$$

- Cost function (want to minimize)

$$J(\boldsymbol{w}) = -\sum_{i=1}^{n} y_i \log h_{\boldsymbol{w}}(\boldsymbol{x_i}) + (1 - y_i)\log(1 - h_{\boldsymbol{w}}(\boldsymbol{x_i}))$$

3 important pieces to SGD

- Hypothesis function (prediction)

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}\cdot\boldsymbol{x}}}$$

- Cost function (want to minimize)

$$J(\boldsymbol{w}) = -\sum_{i=1}^{n} y_i \log h_{\boldsymbol{w}}(\boldsymbol{x_i}) + (1 - y_i)\log(1 - h_{\boldsymbol{w}}(\boldsymbol{x_i}))$$

- Gradient of cost wrt single data point x_i

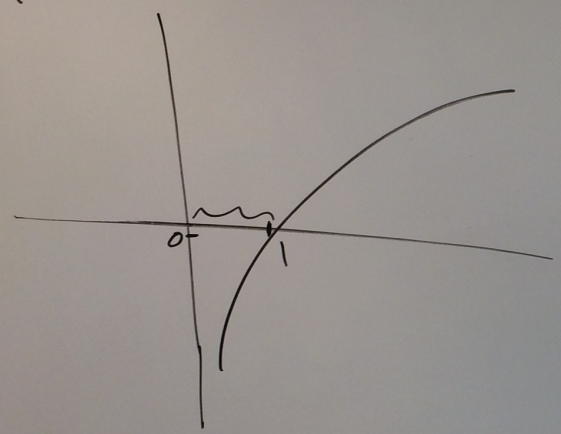
$$\nabla J_{\boldsymbol{x}_i}(\boldsymbol{w}) = (h_{\boldsymbol{w}}(\boldsymbol{x_i}) - y_i)\boldsymbol{x_i}$$

cost = negative log likelihood

$$J(\vec{w}) = -\sum_{i=1}^{n} y_i \log \underbrace{h_{\vec{w}}(\vec{x}_i)}_{\hat{y}_i} + (1-y_i)\log \underbrace{(1-h_{\vec{w}}(\vec{x}_i))}_{\text{prediction of } 0}$$

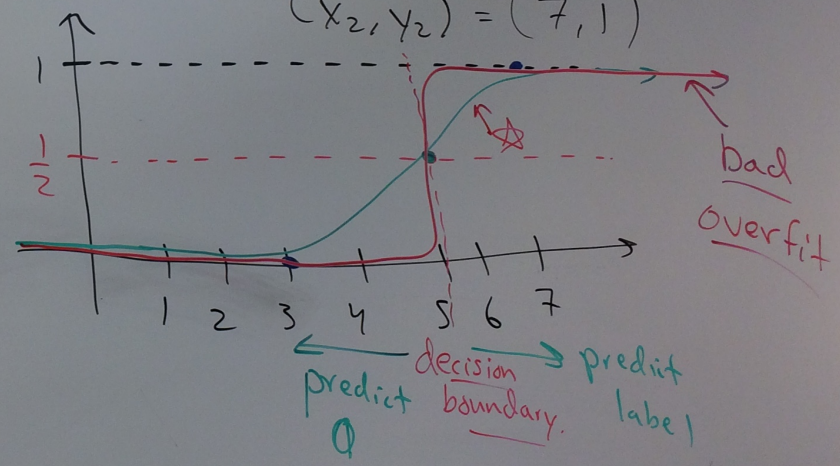
true label

prediction of 1



$(x_1, y_1) = (3, 0)$ &

$(x_2, y_2) = (7, 1)$



if $x > 5 \Rightarrow$ label $\hat{y} = 1$

$x \leq 5 \Rightarrow$ label $\hat{y} = 0$

$a[-5 + x > 0]$

positive constant

$$\boxed{-5a + ax > 0}$$

family of solutions!

$$\boxed{\begin{aligned} \hat{w}_0 &= -5a \\ \hat{w}_1 &= a \end{aligned}}$$

$\lambda =$

Outline for October 10

- Recap SGD for logistic regression
- Regularization
- Multi-class logistic regression
- Begin: evaluation metrics

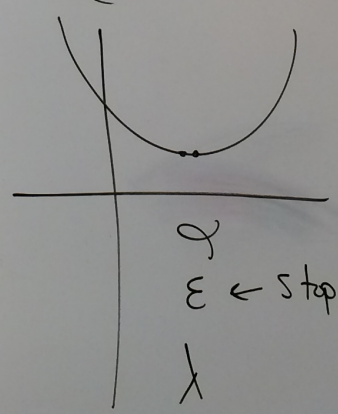
$= 1$

$\hat{y} = 0$

Regularization

$$J^R(w) = \left[-\sum_{i=1}^{n} y_i \log h_{\vec{w}}(\vec{x}_i) + (1-y_i)\log\left(1-h_{\vec{w}}(\vec{x}_i)\right) \right] + \underbrace{\frac{\lambda}{2}\sum_{j=1}^{p} w_j^2}_{\text{don't regularize bias}}$$

λ = regularization parameter
(hyperparameter)
(small & positive)



gradient

$$\nabla J^R_{\vec{x}_i}(\vec{w}) = \left(h_{\vec{w}}(\vec{x}_i) - y_i\right)\vec{x}_i + \lambda \vec{w}^*$$

$$\vec{w}^* = \begin{bmatrix} 0 \\ w_1 \\ w_2 \\ \vdots \\ w_p \end{bmatrix}$$

SGD (regularization)

$$\vec{w} \leftarrow \vec{w} - \alpha\left(\left(h_{\vec{w}}(\vec{x}_i) - y_i\right)\vec{x}_i + \lambda \vec{w}^*\right)$$

$$\leftarrow \underbrace{(1 - \alpha\lambda)}_{\text{less than 1 (pos)}} \vec{w} - \cdots$$

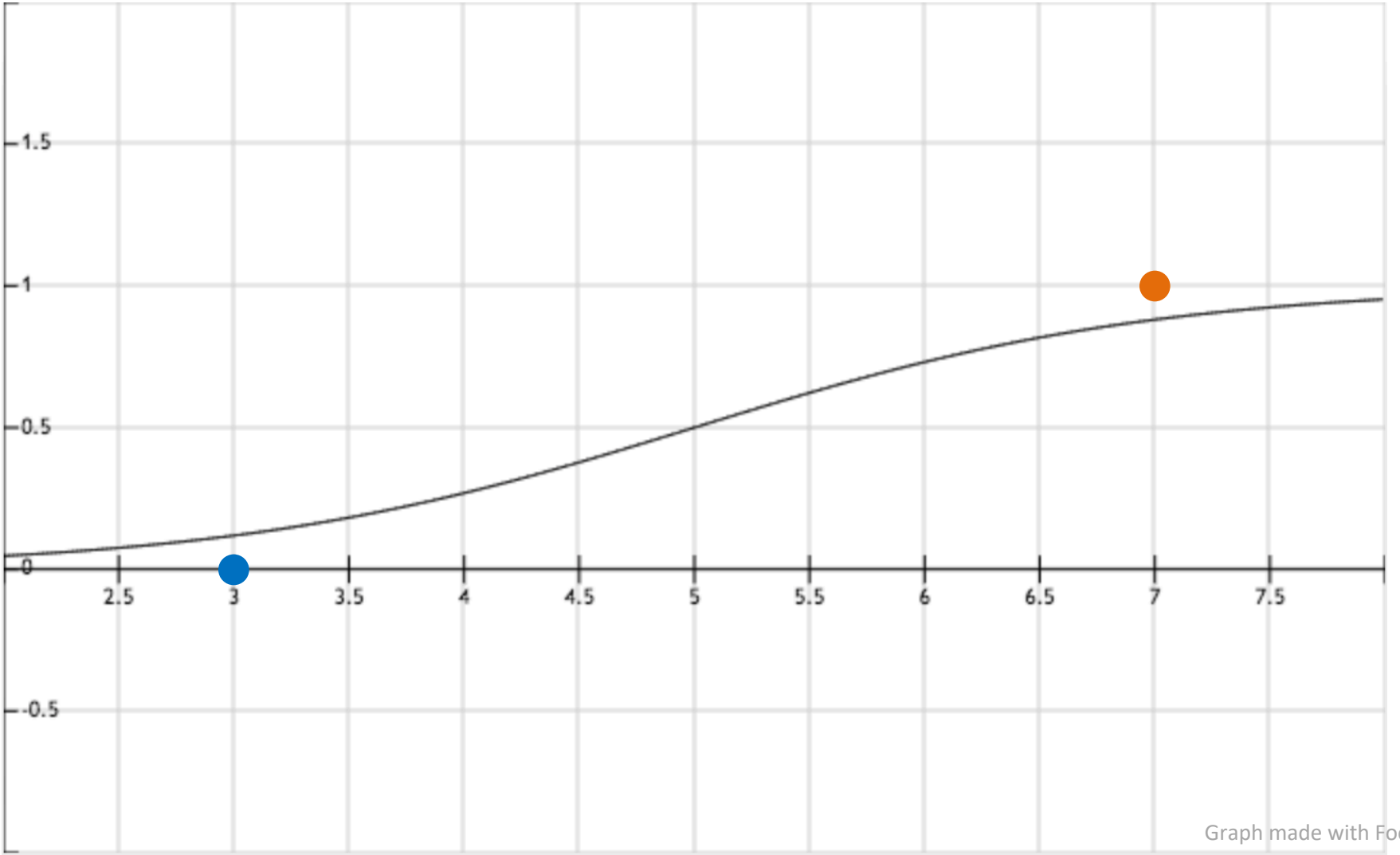
make magnitude of w's smaller

all non-bias terms.

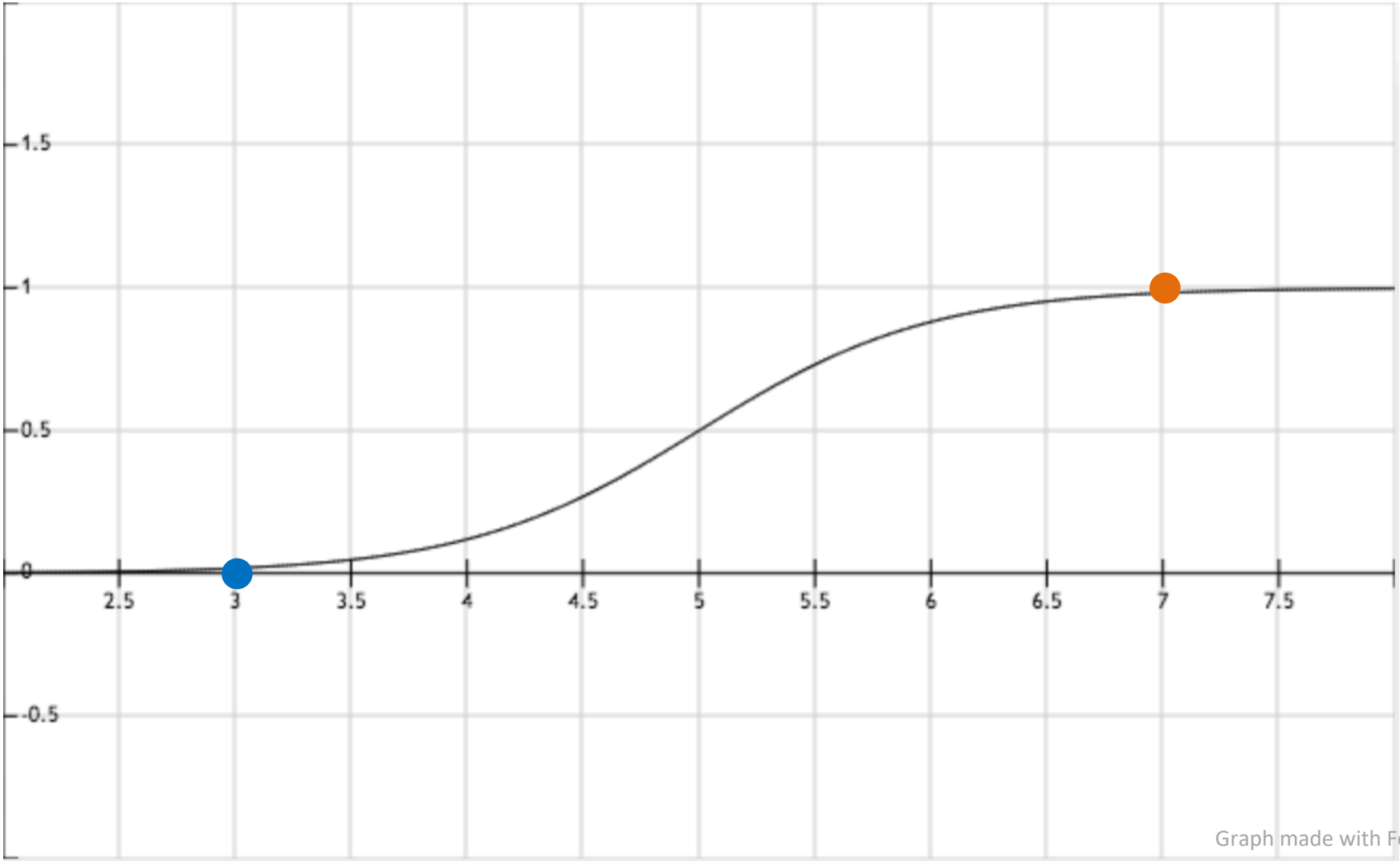
bias

$$w_0 \leftarrow w_0 - \alpha\left(\overset{\text{gradient}}{\cdots}\right)$$

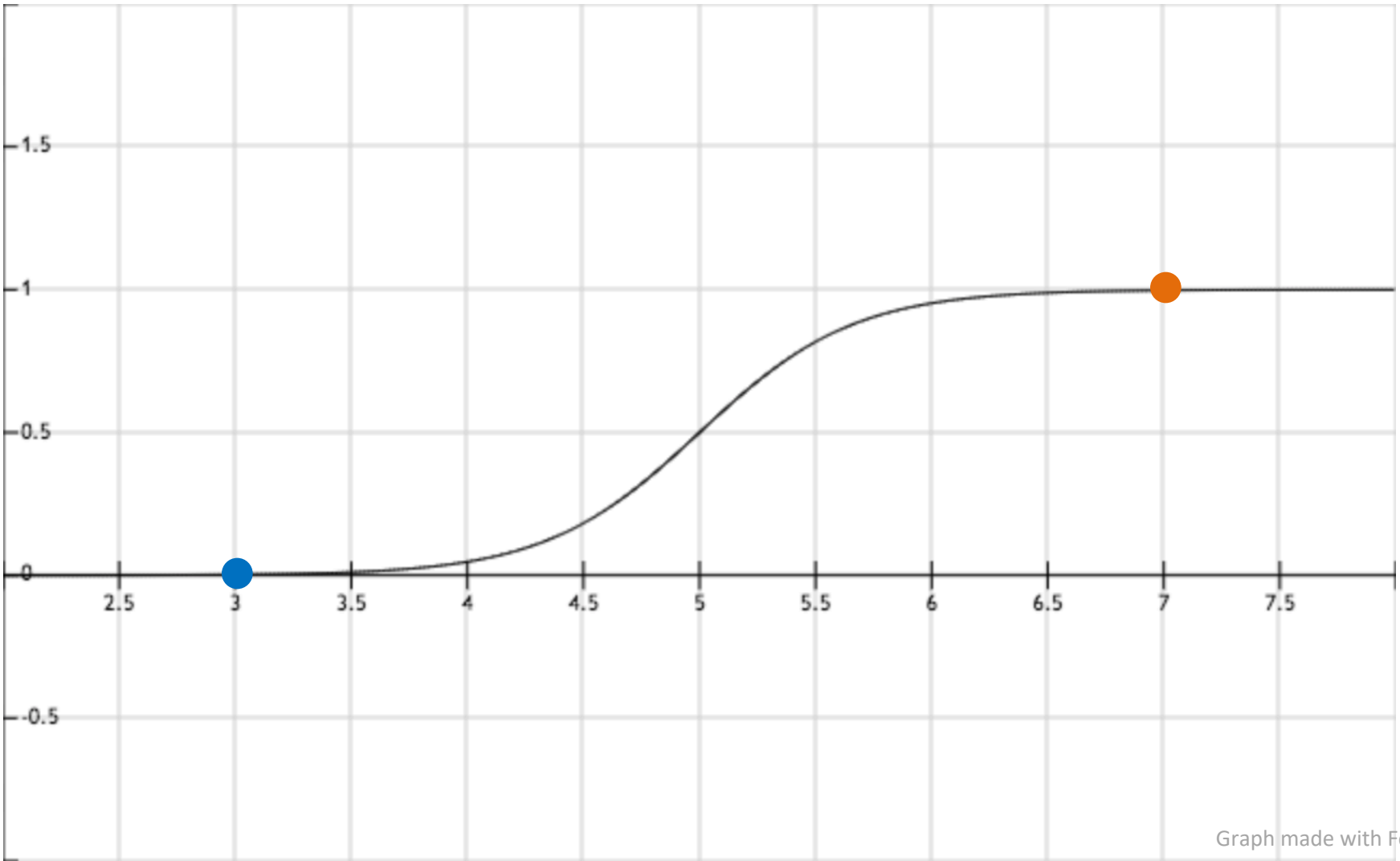
w_0=-5, w_1=1 $h_w(x) = 1/(1+e^{(5-x)})$



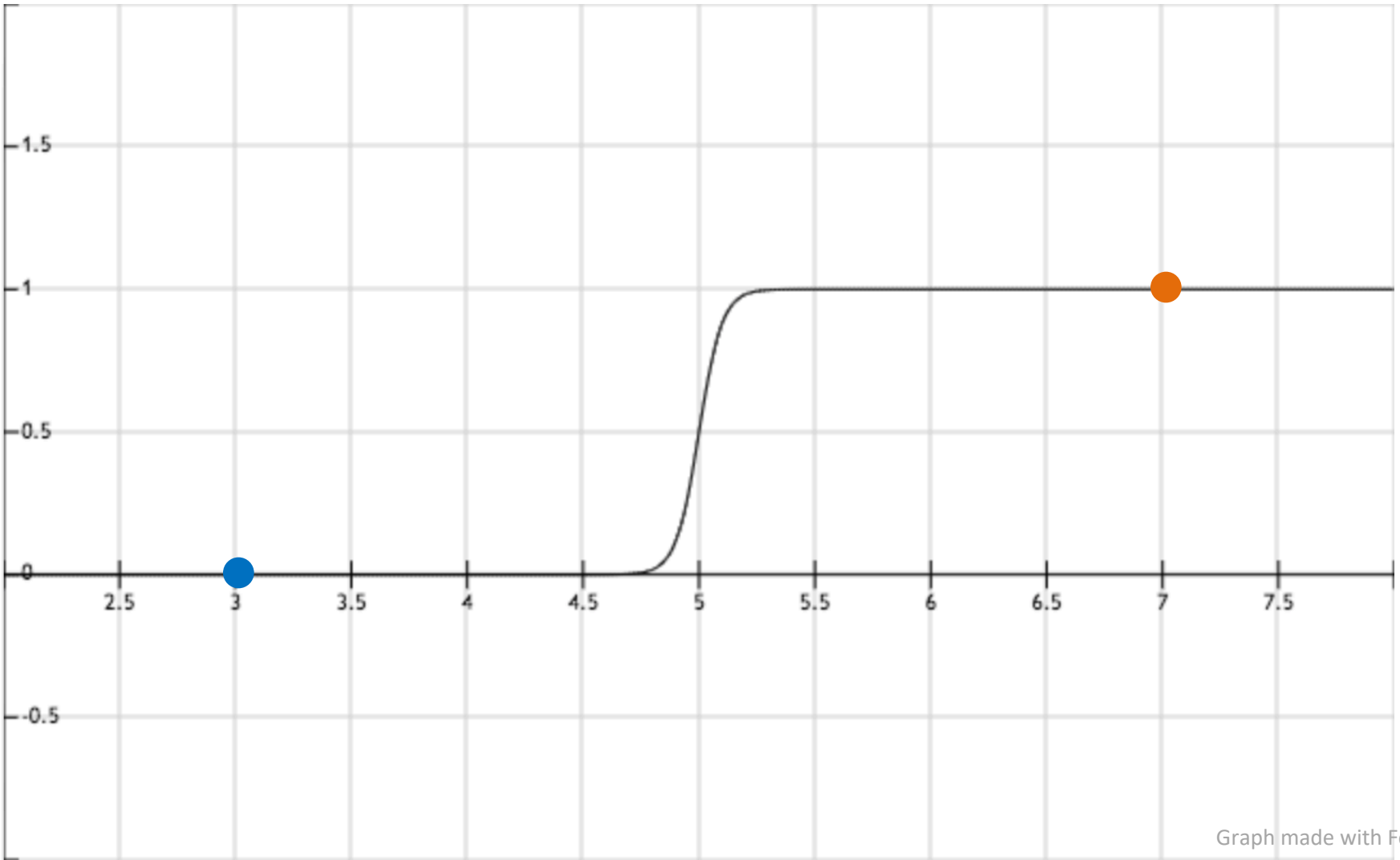
w_0=-10, w_1=2 $h_w(x) = 1/(1+e^{(10-2x)})$



w_0=-15, w_1=3 $h_w(x) = 1/(1+e^{(15-3x)})$



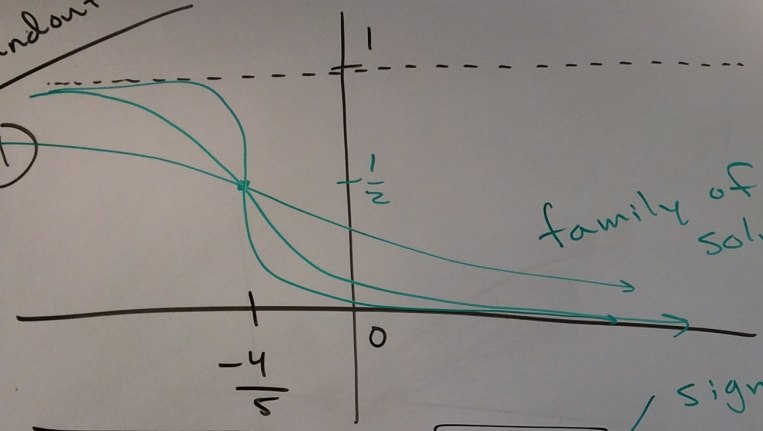
w_0=-100, w_1=20 $h_w(x)$ = 1/(1+e^(100-20x))



Graph made with FooPlot

Handout 11

(1)



$$\boxed{\begin{aligned}\hat{w}_0 &= -4 \cdot a \\ \hat{w}_1 &= -5 \cdot a\end{aligned}}$$

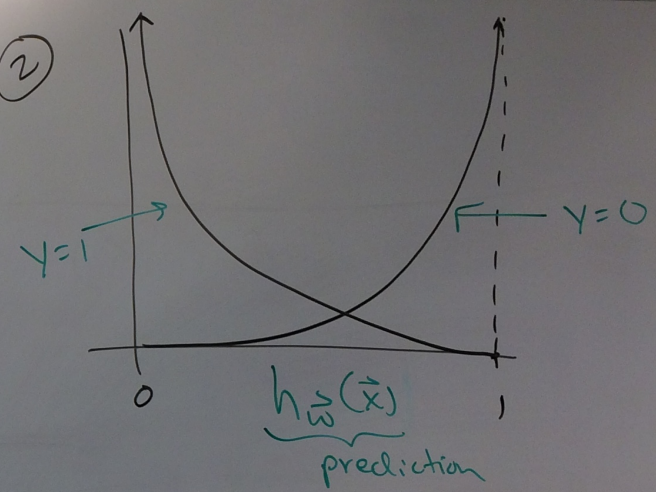
sign of w_1 $\Rightarrow$ slope

$$\boxed{-4 - 5x > 0} \Big\} \text{ pred } 1$$

$$-5x > 4$$

$$\boxed{x < -\frac{4}{5}} \Rightarrow \text{pred } 1$$

(2)



$$h_{\vec{w}}(\vec{x}) = \frac{1}{1+e^{-\vec{w}\cdot\vec{x}}} > \frac{1}{2}$$

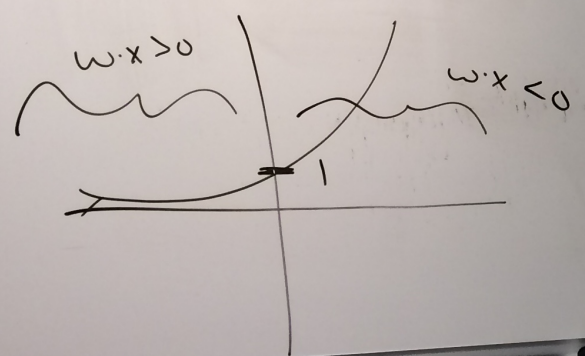
Scalar

$$2 > 1 + e^{-\vec{w}\cdot\vec{x}}$$

$$\log(1) > \log\left(e^{-\vec{w}\cdot\vec{x}}\right)$$

$$0 > -\vec{w}\cdot\vec{x}$$

$$\boxed{\vec{w}\cdot\vec{x} > 0} \Rightarrow \text{predict } 1$$



#3 $g'(z) = -1\,(\text{inside})^{-2}\,(\text{derivative of inside})$

$$= \frac{e^{-z}}{(1+e^{-z})^2}$$

$$= \left(\frac{1}{1+e^{-z}}\right)\left(\frac{1+e^{-z}-1}{1+e^{-z}}\right)$$

$$= g(z)\left(1-g(z)\right)$$

(4) (b) $\begin{bmatrix} 0.05 \\ 0.35 \end{bmatrix}$

(c) $\begin{bmatrix} -0.025 \\ 0.125 \end{bmatrix}$

Outline for October 10

- Recap SGD for logistic regression
- Regularization
- Multi-class logistic regression
- Begin: evaluation metrics

Multiclass Logistic Regression

K classes (political party, blood groups)

2 classes: $h_{\vec{w}}(\vec{x}) = \dfrac{1}{1+e^{-\vec{w}\cdot\vec{x}}} = \dfrac{e^{\vec{w}\cdot\vec{x}}}{\underbrace{e^{\vec{w}\cdot\vec{x}}}_{\text{weight on class 1}} + \underbrace{1}_{\text{weight on class 0}}}$ $\quad 1 = e^0$

K classes

$$\hat{\vec{Y}} = h_w(\vec{x}) = \begin{bmatrix} P(y=1|\vec{x}) \\ P(y=2|\vec{x}) \\ \vdots \\ P(y=K|\vec{x}) \end{bmatrix} = \frac{1}{\sum_{k=1}^{K} e^{\vec{w}^{(k)}\cdot\vec{x}}} \begin{bmatrix} e^{\vec{w}^{(1)}\cdot\vec{x}} \\ e^{\vec{w}^{(2)}\cdot\vec{x}} \\ \vdots \\ e^{\vec{w}^{(K)}\cdot\vec{x}} \end{bmatrix}$$

sum to 1

Weights are a matrix

$$W = \begin{bmatrix} | & | & & | \\ \vec{w}^{(1)} & \vec{w}^{(2)} & \cdots & \vec{w}^{(K)} \\ | & | & & | \end{bmatrix}$$

$(p+1) \times K$

cost

$$J(W) = \underbrace{-\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik} \log \underbrace{p(y_i = k \mid \vec{x}_i)}_{\hat{y}_{ik}}}_{\text{cross entropy}}$$

$$\vec{y}_i = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow \text{index of true label } k$$

y_{i4}

$y_i = 4$

$\rightarrow$

$K = 5$

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \leftarrow y_{i1}$$