CS540 Introduction to Artificial Intelligence
Lecture 2

Young Wu
Based on lecture slides by Jerry Zhu and Yingyu Liang

May 22, 2019

Admin

Admin

- Math and Stat Review posted under W1.
- Complete slides (with diagrams and quiz questions etc) will be posted Thursday or Friday.
- Homework will be posted on Friday (due in 9 days, not 2 days).
- Exact due dates are on Canvas: programming homework can be submitted two weeks late (except for the last two homework (one week late)).
- Lecture this Friday. no lecture Monday

Activation Function

Review

- The supervisered learning problem with activation function is the following.

$$\left(\hat{w}_0, \hat{w}_1, \ldots, \hat{w}_m, \hat{b}\right) = \arg \min_{w_1, \ldots, w_m, b} C$$

$$\text{where } C = \frac{1}{2} \sum_{i=1}^{n} (a_i - y_i)^2$$

$$\text{and } a_i = g\left(w_1 x_{i1} + w_2 x_{i2} + \ldots + w_m x_{im} + b\right)$$

$g = \mathbb{1}_{\{w^T x \geq 0\}}$ ← Perceptron Algorith

Sigmoid Activation Function

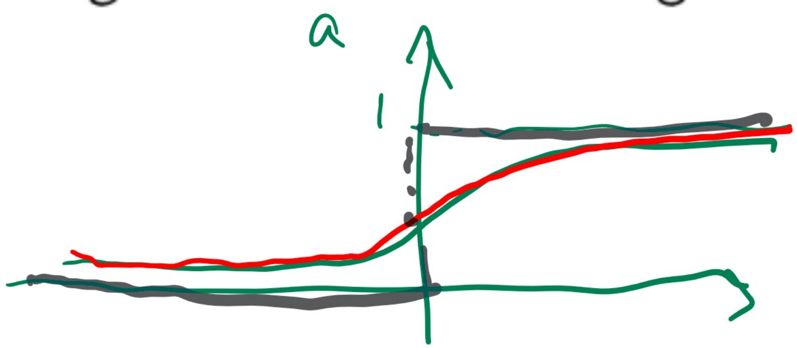
Motivation

- When the activation function g is the sigmoid function, the problem is called logistic regression.

$$g\left(\boxed{\cdot}\right) = \frac{1}{1 + \exp\left(-\boxed{\cdot}\right)}$$

$\mathbb{1}\{\boxed{\cdot} \geq 0\}$

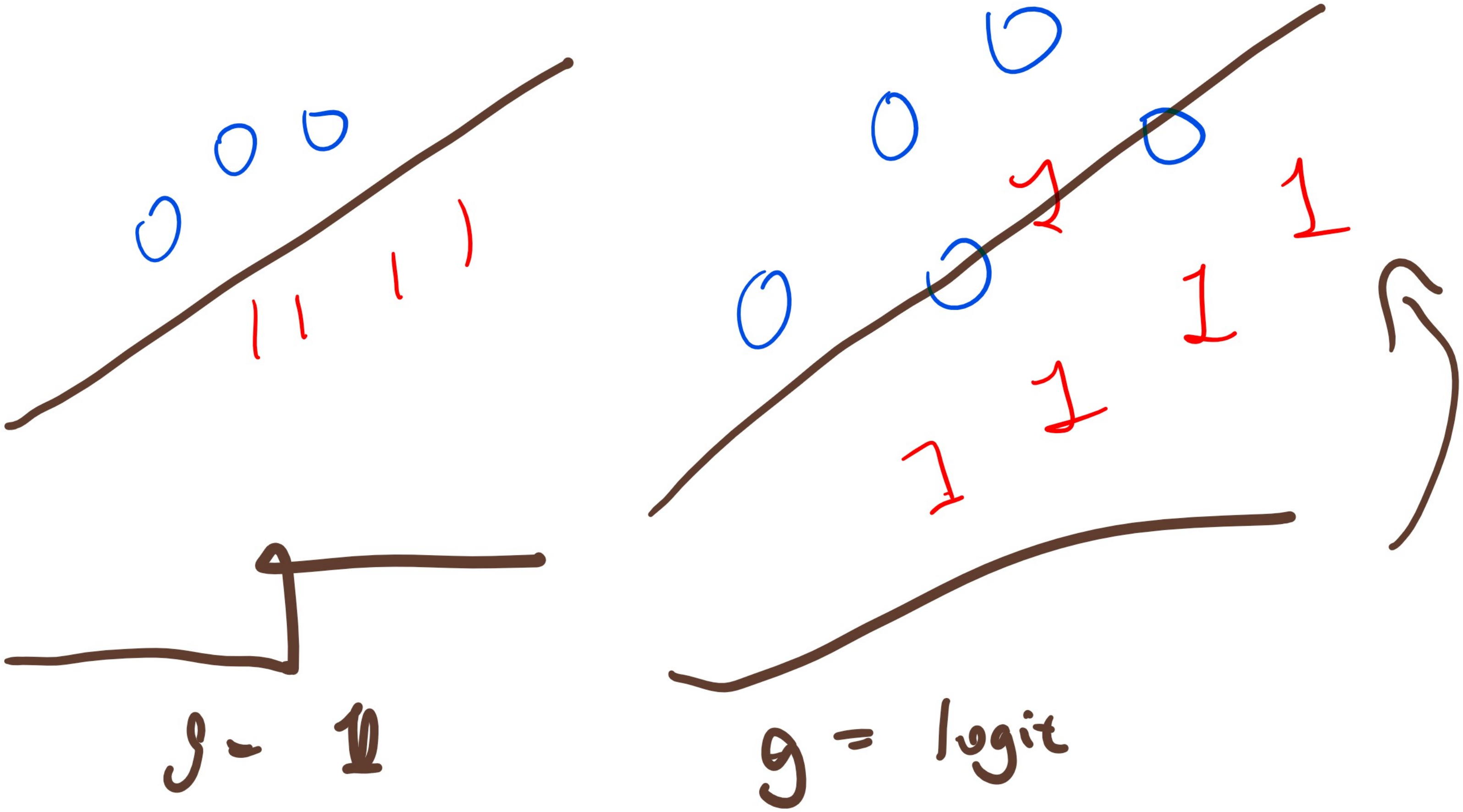
- This g is also called the logistic function.



$\boxed{\cdot} \to w^T x + b$

Sigmoid Function Diagram

Motivation



Cross Entropy Loss Function

Motivation

- The cost function used for logistic regression is usually the log cost function.

$$C = -\sum_{i=1}^{n} \left(y_i \log\left(f\left(x_i\right)\right) + \left(1 - y_i\right)\log\left(1 - f\left(x_i\right)\right)\right)$$

- It is also called the cross-entropy loss function.

Logistic Regression

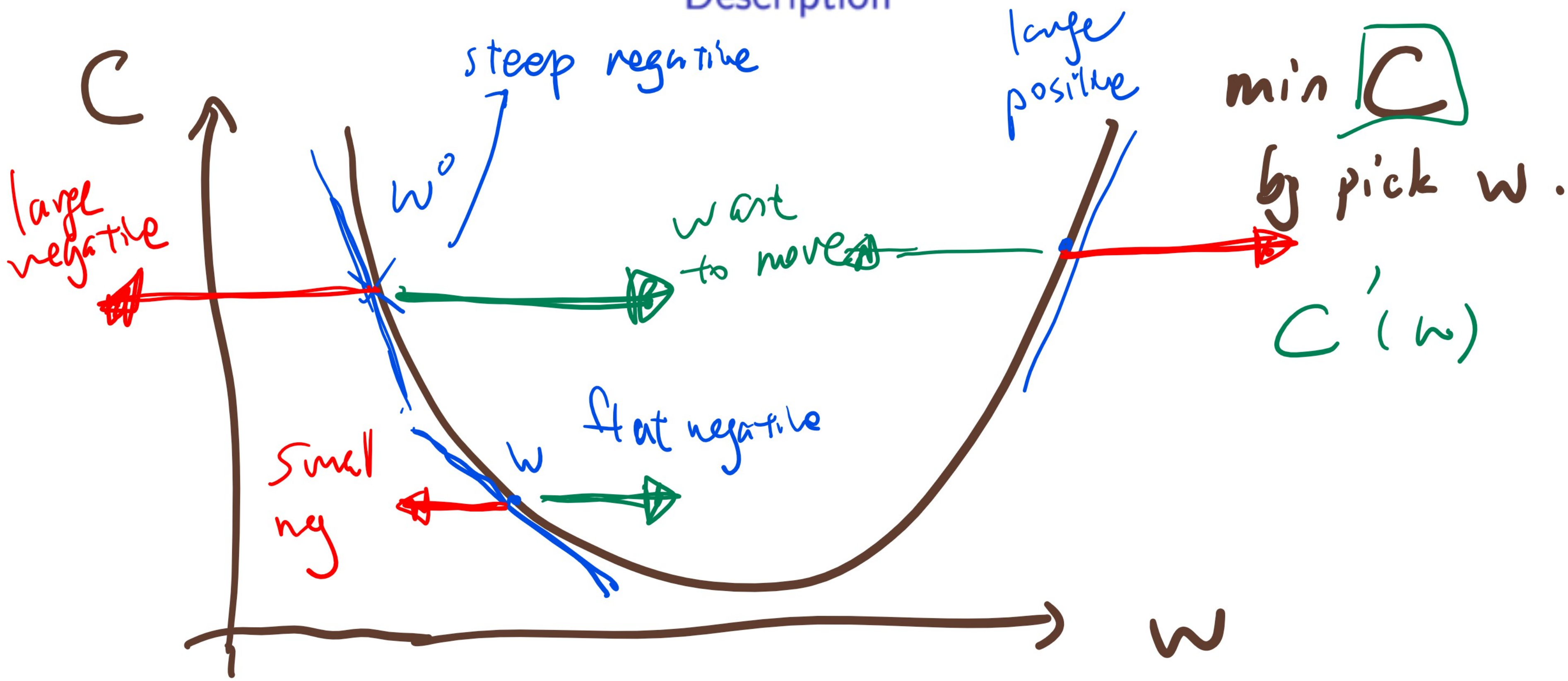
Description

- Initialize random weights.
- Evaluate the activation function.
- Compute the gradient of the cost function with respect to each weight and bias.
- Update the weights and biases using gradient descent.
- Repeat until convergent.

last time → $+ \alpha (a_i - y_i) x_i$

Optimization Diagram

Description



Gradient Descent Intuition

Definition

- If a small increase in w_1 causes the distances from the points to the regression line to decrease: increase $w_{1.}$
- If a small increase in w_1 causes the distances from the points to the regression line to increase: decrease $w_{1.}$
- The change in distance due to change in w_1 is the derivative.
- The change in distance due to change in $\begin{bmatrix} w \\ b \end{bmatrix}$ is the gradient.

Gradient

Definition

- The gradient is the vector of derivatives.
- The gradient of $f\left(x_i\right) = w^T x_i + b = w_1 x_{i1} + w_2 x_{i1} + \ldots + w_m x_{im} + b$ is:

partial derivative

$$\nabla_w f = \begin{bmatrix} \frac{\partial f}{\partial w_1} \\ \frac{\partial f}{\partial w_2} \\ \ldots \\ \frac{\partial f}{\partial w_m} \end{bmatrix} = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \ldots \\ x_{im} \end{bmatrix} = x_i$$

$\nabla_w w^T x_i = x_i$

$$\nabla_b f = 1$$

Chain Rule

Definition

- The gradient of
  $f\left(x_i\right) = g\left(w^T x_i + b\right) = g\left(w_1 x_{i1} + w_2 x_{i1} + \ldots + w_m x_{im} + b\right)$
  can be found using the chain rule.

$\nabla_w \left(w^T x_i + b\right)$

$$\nabla_w f = g'\left(w^T x_i + b\right) x_i$$
$$\nabla_b f = g'\left(w^T x_i + b\right) \cdot 1$$

- In particular, for the logistic function g :

$$g\left(\boxed{\cdot}\right) = \frac{1}{1 + \exp\left(-\boxed{\cdot}\right)}$$
$$g'\left(\boxed{\cdot}\right) = g\left(\boxed{\cdot}\right)\left(1 - g\left(\boxed{\cdot}\right)\right)$$

Logistic Gradient Derivation

Definition

$$g(x) = \frac{1}{1+e^{-x}} = (1+e^{-x})^{-1}$$

$$\frac{d}{dx}(1+e^{-x})$$

$$g'(x) = -1 \cdot (1+e^{-x})^{-2} \cdot \boxed{e^{-x} \cdot (-1)}$$

$$= \frac{1}{1+e^{-x}} \frac{e^{-x}}{1+e^{-x}}$$

$$= g(x)\,(1-g(x))$$

Gradient Descent Step

Definition

- For logistic regression, use chain rule twice.

$$w = w - \alpha \sum_{i=1}^{n} \left(a_i - y_i\right) x_i$$

$$b = b - \alpha \sum_{i=1}^{n} \left(a_i - y_i\right)$$

$$a_i = g\left(w^T x_i + b\right), g\left(\boxed{\cdot}\right) = \frac{1}{1 + \exp\left(-\boxed{\cdot}\right)}$$

- α is the learning rate. It is the step size for each step of gradient descent.

Gradient Descent Derivation

Definition

$$\nabla_w C = \begin{bmatrix} \frac{\partial C}{\partial w_1} \\ \vdots \\ \vdots \end{bmatrix}$$

Chain Rule

$$\frac{\partial C}{\partial w_j} = \sum_{i=1}^{n} \frac{\partial C}{\partial a_i} \frac{\partial a_i}{\partial w_j}$$

$$\frac{\partial g}{\partial w_j} = \boxed{a_i(1-a_i)x_{ij}}$$

from prev slide

$$\boxed{C = -\sum_i y_i \log a_i + (1-y_i)\log(1-a_i)}$$

$$\frac{dC}{da_i} = -\frac{y_i}{a_i} - \frac{1-y_i}{1-a_i}(-1) = -\frac{y_i}{a_i} + \frac{1-y_i}{1-a_i}$$

$$\frac{\partial C}{\partial w_j} = \left(-\frac{y_i}{a_i} + \frac{1-y_i}{1-a_i}\right) a_i (1-a_i) x_{ij}$$

Learning Rate Diagram

Definition

$$\Big(-(1-a_i)y_i + a_i(1-y_i)\Big)x_{ij}$$

$$\frac{\partial C}{\partial w_j} = \sum_{i=1}^{n}(a_i - y_i)x_{ij}$$

$$\nabla_w C = \sum_{i=1}^{n}(a_i - y_i)x_i$$

gradient vector

$$\begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{im} \end{pmatrix}$$

$$w = w - \alpha \nabla_w C$$

3

Gradient Descent

Quiz (Graded)

ignore ①

- What is the gradient descent step for w if the objective (cost) function is the squared error?

$$C \simeq \cancel{\frac{1}{2}} \sum_{i=1}^{n} (a_i - y_i)^{\cancel{2}}$$

$$\boxed{\nabla_w a_i = a_i(1-a_i)x_i}$$

- A: $w = w - \alpha \sum (a_i - y_i)\, x_i$
- B: $w = w - \alpha \sum (a_i - y_i)\, a_i x_i$
- C: $w = w - \alpha \sum (a_i - y_i)\, (1 - a_i)\, x_i$
- ✓ D: $w = w - \alpha \sum \boxed{(a_i - y_i)\, a_i\, (1 - a_i)\, x_i}$
- E: None of the above

$$\nabla_w C = \frac{\partial C}{\partial a_i} \nabla_w a_i$$

$$= (a_i - y_i)$$

$$|x| = \begin{cases} x & \text{if } x>0 \\ -x & \text{if } x<0 \end{cases}$$

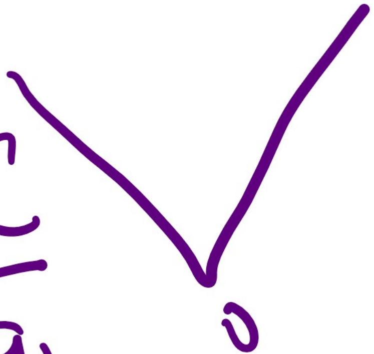
Gradient Descent, Another One

Quiz (Graded)

- What is the gradient descent step for w if the objective (cost) function is the absolute value error?

$$C = \sum_{i=1}^{n} |a_i - y_i|$$

$$\frac{\partial C}{\partial a_i}$$



→ for logit

$$\nabla_w a_i = \sum_{i=1}^{n} a_i(1-a_i)x_i$$

- A: $w = w - \alpha \sum (a_i - y_i)\, a_i\,(1 - a_i)\, x_i$
- B: $w = w - \alpha \sum |a_i - y_i|\, a_i\,(1 - a_i)\, x_i$
- C: $w = w - \alpha \sum \mathbb{1}_{\{a_i - y_i > 0\}} a_i\,(1 - a_i)\, x_i$

$$\begin{cases} 1 & a_i - y_i > 0 \\ 0 & \text{not} \end{cases}$$

- D: $w = w - \alpha \sum \boxed{\text{sign}\,(a_i - y_i)}\, a_i\,(1 - a_i)\, x_i$
- E: None of the above

$$\begin{cases} +1 & a_i - y_i > 0 \\ 0 & a_i = y_i \\ -1 & a_i - y_i < 0 \end{cases}$$

Logistic Regression, Part 1

Algorithm

- Inputs: instances: $\{x_i\}_{i=1}^{n}$ and $\{y_i\}_{i=1}^{n}$
- Outputs: weights and biases: $w_1, w_2, ..., w_m$ and b
- Initialize the weights.

$$w_1, ..., w_m, b \sim \text{ Unif } [0, 1]$$

$$\boxed{w = \left(X^T X\right)^{-1} X^T y}$$

$$g(x) = x$$

- Evaluate the activation function.

$$a_i = g\left(w^T x_i\right), g\left(\boxed{\cdot}\right) = \frac{1}{1 + \exp\left(-\boxed{\cdot}\right)}$$

Logistic Regression, Part 2

Algorithm

- Update the weights and bias using gradient descent.

∇

$-\alpha\nabla$

$$w = w - \alpha \sum_{i=1}^{n} (a_i - y_i)\, x_i$$

$$b = b - \alpha \sum_{i=1}^{n} (a_i - y_i)$$

- Repeat the process until convergent.

$$|C - C^{\text{prev}}| < \varepsilon \approx 0.0001$$

Other Non-linear Activation Function

Discussion

- Activation function: $g\left(\boxed{\cdot}\right) = \tanh\left(\boxed{\cdot}\right) = \dfrac{e^{\boxed{\cdot}} - e^{\boxed{\cdot}}}{e^{\boxed{\cdot}} + e^{\boxed{\cdot}}}$
- Activation function: $g\left(\boxed{\cdot}\right) = \arctan\left(\boxed{\cdot}\right)$
- Activation function (rectified linear unit): $g\left(\boxed{\cdot}\right) = \boxed{\cdot}\,\mathbb{1}_{\left\{\boxed{\cdot}\geqslant 0\right\}}$
- All these functions lead to objective functions that are convex and differentiable. Gradient descent can be used.

Convexity Diagram

Discussion

logistic C

Convexity

Discussion

- If a function is convex, gradient descent with any initialization will converge to the global minimum.
- If a function is not convex, gradient descent with different initializations may converge to different local minima.
- A twice differentiable function is convex if and only its second derivative is non-negative.
- In the multivariate case, it means the Hessian matrix is positive semidefinite.

Positive Semidefinite

Discussion

- Hessian matrix is the matrix of second derivatives:

$$H : H_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

$f(x_1, x_2)$

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix}$$

- A matrix H is positive semidefinite if $x^T H x \geqslant 0 \; \forall \; x \in \mathbb{R}^n$.
- A symmetric matrix is positive semidefinite if and only if all of its eigenvalues are non-negative.

$$H x = \underbrace{\lambda}_{\text{eigenvalue}} \overbrace{x}^{\text{eigenvector}}$$

Convex Functions

Quiz (Participation)

- What is the Hessian (second derivative) of

$$f(x) = x_1^2 + 4x_1x_2 + x_2^2 = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$2x_1 + 4x_2$

2

$2x_1 + 4x_2$

4

$$\begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \partial x_2} \\ 0 & \dfrac{\partial^2 f}{\partial x_2^2} \end{bmatrix}$$

- A: $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$
- B: Do not choose this.
- C: $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$
- D: Do not choose this.
- E: $\begin{bmatrix} 1 & 4 \\ 4 & 1 \end{bmatrix}$

$$\begin{matrix} 2 & 4 \\ 4 & 2 \end{matrix}$$

Definiteness

Quiz (Participation)

put on midterm.

$\lambda = -1$

- Which ones (two) of the following are the eigenvalues of $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$? Two eigenvectors are $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$.
- A: 0
- B: 1
- C: 2
- D: 3 ✓ 3
- E: 4

$Hx = \lambda x$

e.v.

$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

$\begin{bmatrix} 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = ? \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

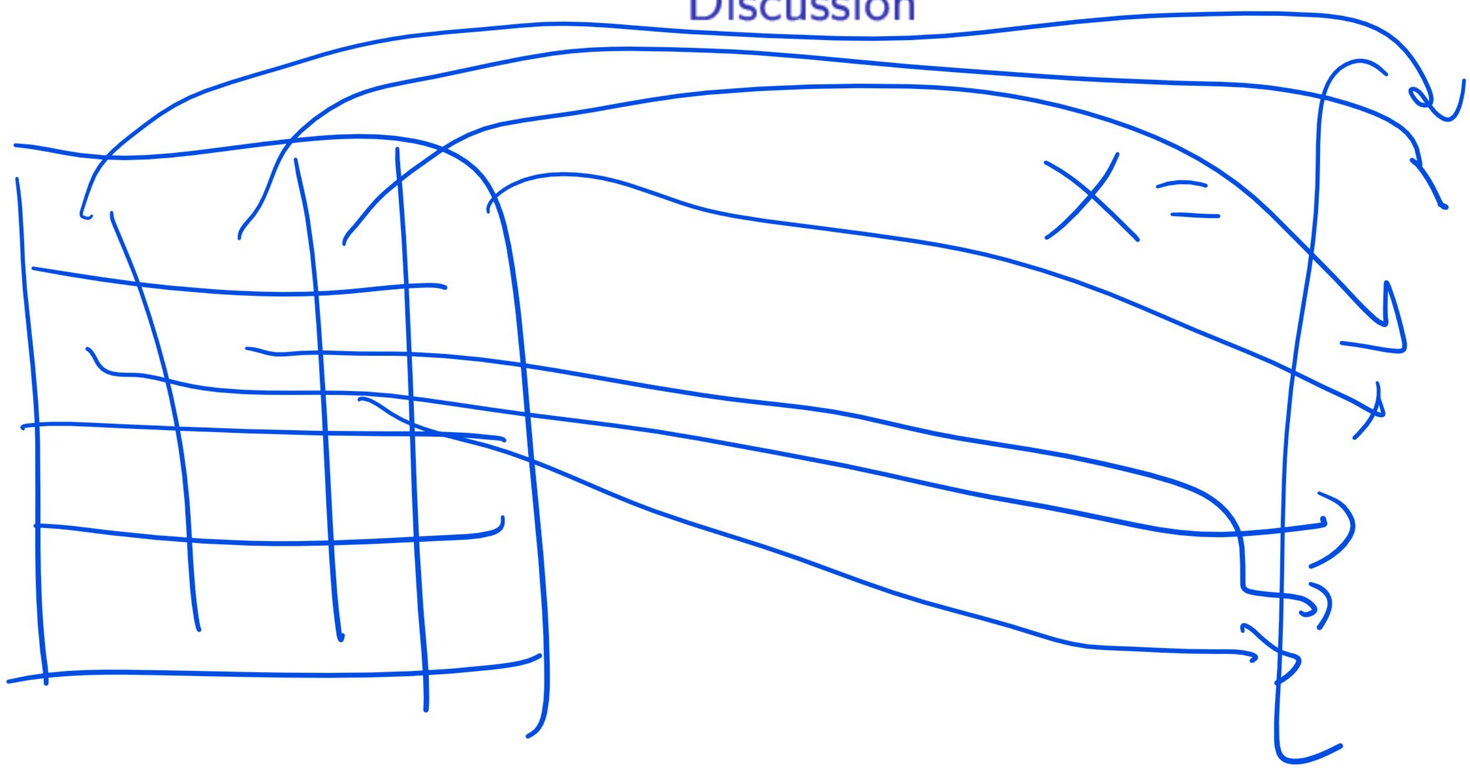
Image as Input

Discussion

- Simplest feature vector for an image is the flattened pixel intensities.
- One way to compute pixel intensity is to use the average of the RGB values divided by 255.
- Pixel intensity of each pixel is between 0 and 1.
- An n_w pixel by n_h pixel image then can be flattened into a $m = n_w n_h$ dimensional input feature vector x.

Flattened Feature Vector Diagram

Discussion



$g(w^T x)$

AND Operator Data

Quiz (Particpation)

- Sample data for AND

x_1	x_2	y
0	0	0
0	1	0
1	0	0
1	1	1

Learning AND Operator

Quiz (Particpation)

- Which one of the following is AND?
- A: $\hat{y} = \mathbb{1}_{\{1x_1 + 1x_2 - 1.5 \geqslant 0\}}$
- B: $\hat{y} = \mathbb{1}_{\{1x_1 + 1x_2 - 0.5 \geqslant 0\}}$
- C: $\hat{y} = \mathbb{1}_{\{-1x_1 + 0.5 \geqslant 0\}}$
- D: $\hat{y} = \mathbb{1}_{\{-1x_1 - 1x_2 + 0.5 \geqslant 0\}}$
- E: None of the above

OR Operator Data

Quiz (Graded)

- Sample data for OR

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	1

Learning OR Operator

Quiz (Graded)

- Which one of the following is OR?
- A: $\hat{y} = \mathbb{1}_{\{1x_1+1x_2-1.5 \geqslant 0\}}$
- B: $\hat{y} = \mathbb{1}_{\{1x_1+1x_2-0.5 \geqslant 0\}}$
- C: $\hat{y} = \mathbb{1}_{\{-1x_1+0.5 \geqslant 0\}}$
- D: $\hat{y} = \mathbb{1}_{\{-1x_1-1x_2+0.5 \geqslant 0\}}$
- E: None of the above

XOR Data

Quiz (Graded)

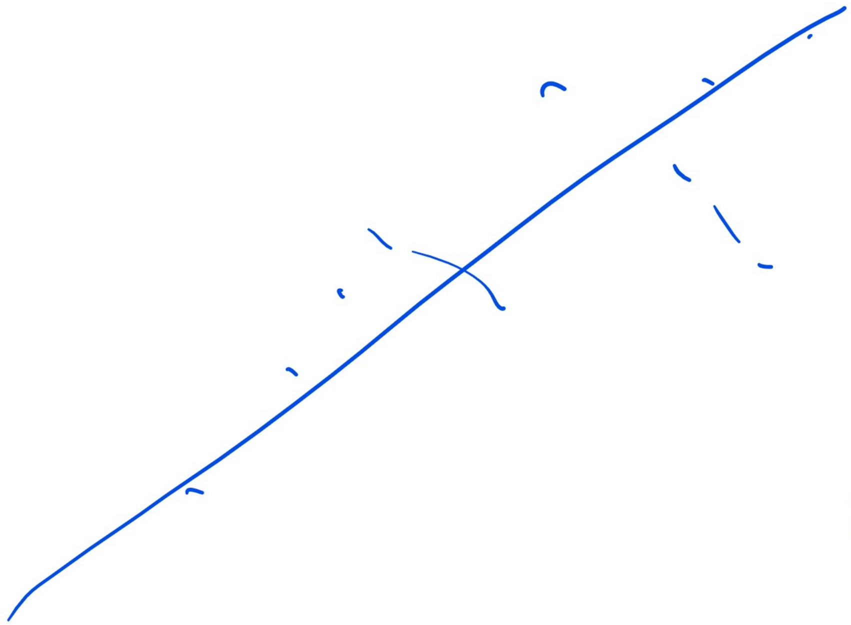
- Sample data for XOR

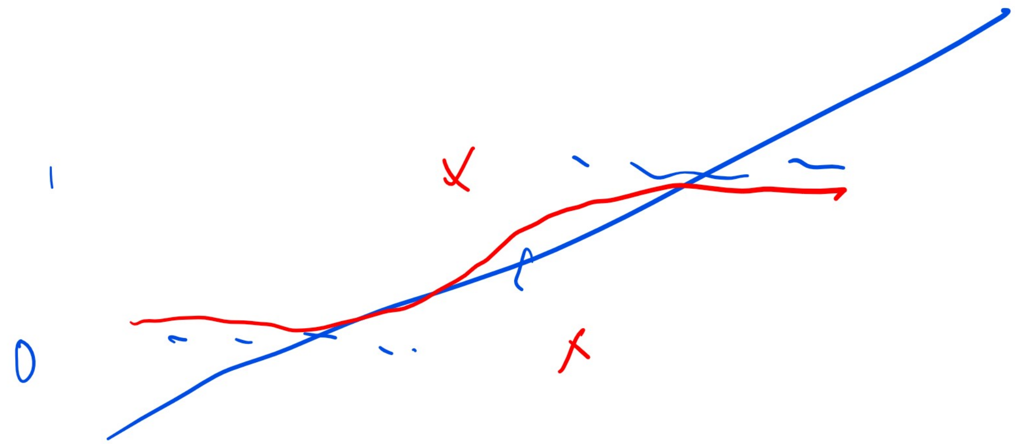
x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

Learning XOR Operator

Quiz (Graded)

- Which one of the following is XOR?
- A: $\hat{y} = \mathbb{1}_{\{1x_1 + 1x_2 - 1.5 \geqslant 0\}}$
- B: $\hat{y} = \mathbb{1}_{\{1x_1 + 1x_2 - 0.5 \geqslant 0\}}$
- C: $\hat{y} = \mathbb{1}_{\{-1x_1 + 0.5 \geqslant 0\}}$
- D: $\hat{y} = \mathbb{1}_{\{-1x_1 - 1x_2 + 0.5 \geqslant 0\}}$
- E: None of the above





$$y = wx + b$$

$$\hat{y} = \begin{cases} 1 & \text{if } y = \underline{w}^T x > 0.5 \\ 0 & \text{———} < 0.5 \end{cases}$$