

Review
o

Logistic
oooooooooooooooooooo

Demo
oo

Multiclass
oooo

Lab
oo

Day 4: Classification

Summer STEM: Machine Learning

Department of Electrical and Computer Engineering
NYU Tandon School of Engineering
Brooklyn, New York

August 6, 2020

Outline

1 Review

2 Logistic Regression

3 Lab: Diagnosing Breast Cancer

4 Multiclass Classification

5 Lab: Iris Dataset

Machine Learning

{ unsupervised

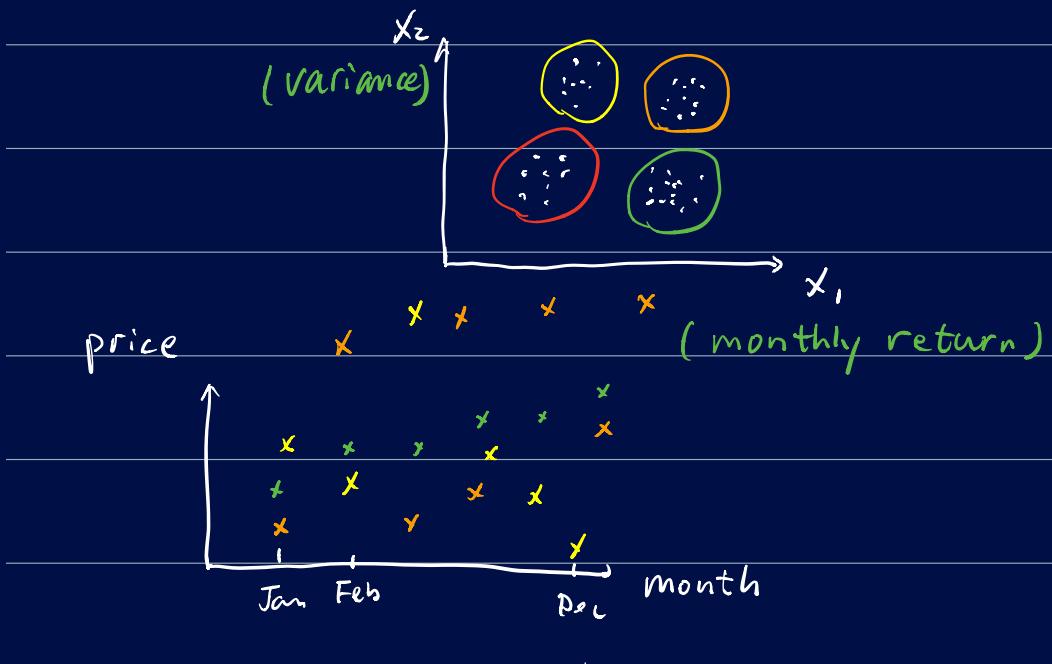
supervised (regression / classification)

- learn patterns / trends from data
- without specific rules / instructions (autonomous)
- predict

Data features, labels

Supervised data with labels

unsupervised learning data without labels



Supervised Learning

- Classification labels: different classes

discrete number

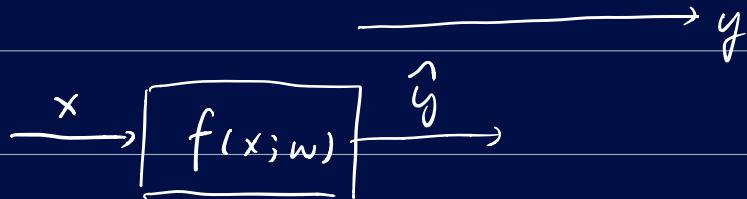
binary { 0, 1)

↑ ↑

Cats dogs

{ 0, 1, 2 } multiclass
↑ ↑ ↑ classification
planes trains cars

- Regression labels: real numbers



$$w: f(x; w) = w_0 + w_1 x \quad w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$\text{Vector-valued features } x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \end{bmatrix}$$

Remark: We use the subscript i in x_i to denote the i -th sample; but sometimes we use the same notation to denote the i -th component of a vector x .

$$\underline{f(x_i; w) = w_0 + w_1 x_{i1} + w_2 x_{i2}}, \quad w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$$

a model parameterized by w

$$f(x)$$

Scalar-valued feature \times Polynomial with
an order of M

$$\underline{a_0x^0 + a_1x^1 + a_2x^2 + a_3x^3 + \dots + a_Mx^M}$$

the order of this polynomial is the highest power

$$a_i x^i \text{ where } a_i \neq 0$$

$$\text{Ex: } x + x^4 \rightarrow 4\text{-th order}$$

| 0-th order

$$x + 3x^4 + 5x^M \quad M\text{-th order if } M \geq 4$$

$$f(x; w) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M \quad (w_M \neq 0)$$

$$w = [w_0 \ w_1 \ \dots \ w_M]^T$$

Matrix-vector representation of the model

$$f(x) = w_0 + w_1 x$$

$$= [1 \ x] \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

let's do this for $i = 1, 2, \dots, N$

$$\hat{y}_1 = [1 \ x_1] \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$\hat{y}_2 = [1 \ x_2] \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \Rightarrow \hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$\hat{y}_N = [1 \ x_N] \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

Vector-valued features

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \end{bmatrix}$$

$$f(x; w) = w_0 + w_1 x_{i1} + w_2 x_{i2}$$

$$= [1 \ x_{i1} \ x_{i2}] \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$$

$$\hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{N1} & x_{N2} \end{bmatrix}}_X \underbrace{\begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}}_w$$

X

The shape of X , w ?

$X \in (N, 3)$

$w \in (3, 1)$

$\hat{Y} \in (N, 1)$

$$\hat{Y} = Xw$$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{N1} & x_{N2} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} w_0 + w_1 x_{11} + w_2 x_{12} \\ w_0 + w_1 x_{21} + w_2 x_{22} \\ \vdots \\ w_0 + w_1 x_{N1} + w_2 x_{N2} \end{bmatrix}$$

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \end{bmatrix} \quad f(x) = w_0 + w_1 x_{i1} + w_2 x_{i2} = \hat{y}_i$$

$$\hat{Y} = Xw \quad w = \underbrace{(X^T X)^{-1} X^T Y}_{\text{f}} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

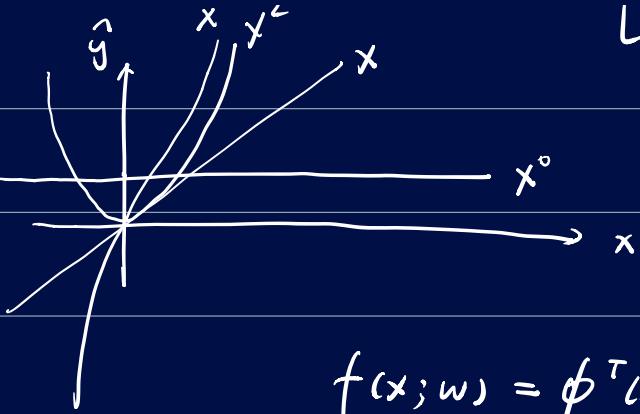
Scalar-valued features x :

$$f(x_i; w) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M$$

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^M \\ 1 & x_2 & x_2^2 & \dots & x_2^M \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^M \end{bmatrix} \quad \underline{Xw}$$

$$\frac{x}{\downarrow} \text{ feature transformation } \phi(x) = [1 \ x \ x^2 \ \dots \ x^M]^T$$

$$\underline{[1 \ x \ x^2 \ \dots \ x^M]^T} \quad \phi_j(x) = x^j \quad \begin{array}{l} \text{basis} \\ \text{function} \end{array}$$



$$\begin{bmatrix} 1 & x & x^2 & \dots & x^n \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix}$$

$$f(x; w) = \phi^T(x) w = \underbrace{w^T \phi(x)}$$

$$(a^T b = b^T a \text{ (a, b vectors)})$$

$$f(x_i; w) = w^T \phi(x_i) \quad M=1$$

scalar-valued $\underbrace{\phi(x_i)}_{=} = [1 \ x_i]^T$

$$\rightarrow \underbrace{\phi(x_i)}_{=} = [1 \ x_i \ x_i^2 \ \dots \ x_i^M]^T$$

vector-valued $\underbrace{\phi(x_i)}_{=} = [1 \ x_{i1} \ x_{i2} \ \dots \ x_{iD}]^T$

$$x_i = [x_{i1} \ x_{i2} \ \dots \ x_{iD}]^T$$

$$f(x_i; w) = \underbrace{w^T \phi(x_i)}_{M=2}$$

$$x_i = [x_{i1} \ x_{i2}]^T$$

$$f(x_i; w) = w_{00} + w_{10} x_{i1} + w_{01} x_{i2}$$

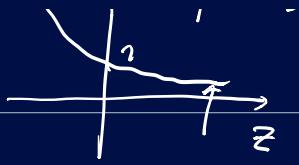
$$+ w_{20} x_{i1}^2 + w_{02} x_{i2}^2 + w_{11} x_{i1} x_{i2}$$

$$= \sum_j \sum_k w_{jk} x_{i1}^j x_{i2}^k$$

$$(j+k \leq 2)$$

\ * exp(-z)

$$\phi_j(x_i) = \exp\left(-\frac{\|x_i - x_j\|^2}{2s^2}\right)$$



Now that we have data, models, how do we evaluate if a model is good or bad?

Loss functions $J(w) = \frac{1}{N} \|Y - \hat{Y}\|^2$ (MSE)

$$= \frac{1}{N} \sum_{i=1}^N \|y_i - \hat{y}_i\|^2$$

for linear models $\hat{Y} = Xw$

$$X = \begin{bmatrix} \phi^\top(x_1) \\ \phi^\top(x_2) \\ \vdots \\ \phi^\top(x_N) \end{bmatrix}$$

(MSE) $J(w) = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$

MSE of 0.7 lbs

MSE of 0.3 lbs

$$RMSE = \sqrt{MSE}$$

$$\underline{w = (X^\top X)^{-1} X^\top Y}$$

Overfitting and Generalization

- Learn patterns/trends from the data to predict the labels for unseen features
- We say a model generalizes well if it has a good performance for unseen data.
- if a model performs well on the training set but poorly on unseen data \rightarrow overfitting

\rightarrow How do we alleviate overfitting?

- Reduce M (Polynomials)
 - weight-based regularization
- $\left. \begin{array}{l} \\ \end{array} \right\} \Rightarrow$ limit the complexity of the model



- Hyper-parameters
- M order of the polynomial
- λ regularization coefficient

$$J(w) = \left| \frac{1}{N} \sum_{i=1}^N \|y_i - w^\top \phi(x_i)\|^2 \right| + \lambda \|w\|_2^2$$

f(x_i; w)

regularization term

MSE

$$\begin{aligned}\lambda \|w\|^2 &= \lambda \left(\sqrt{w_0^2 + w_1^2 + \dots + w_M^2} \right)^2 \\ &= \lambda (w_0^2 + w_1^2 + \dots + w_M^2) \\ &= \lambda \sum_{j=0}^M w_j^2\end{aligned}$$

Norm-based regularization

L2-norm $\|w\|_2 = \sqrt{w_0^2 + w_1^2 + \dots + w_M^2}$

L1-norm $\|w\|_1 = |w_0| + |w_1| + \dots + |w_M|$

$\lambda \|w\|_2^2 \rightarrow$ Ridge regularization

$\lambda \|w\|_1 \rightarrow$ Lasso regularization

- Ridge regularization $\lambda = 0 \quad \lambda \uparrow \quad \|w^*\| \downarrow$

$$w^* = (X_{\text{train}}^\top X_{\text{train}} + \lambda I)^{-1} X_{\text{train}} Y_{\text{train}}$$

- Lasso regularization $\lambda \uparrow \rightarrow$ more components of w become 0

How would it be useful?

$$f(x; w) = w_0 + w_1 x + w_2 x^2 + \dots + w_m x^m$$

Ex $x_i = \begin{matrix} x_{i1} \\ age \end{matrix}, \begin{matrix} x_{i2} \\ height \end{matrix}, \begin{matrix} x_{i3} \\ eye \color{red} \text{color} \end{matrix}$

y_i : weight

$$\hat{y}_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i3}$$

Lasso $w_3 \rightarrow 0$

Review
o

Logistic
●oooooooooooooooooooo

Demo
oo

Multiclass
oooo

Lab
oo

Outline

1 Review

2 Logistic Regression

3 Lab: Diagnosing Breast Cancer

4 Multiclass Classification

5 Lab: Iris Dataset

Classification Vs. Regression

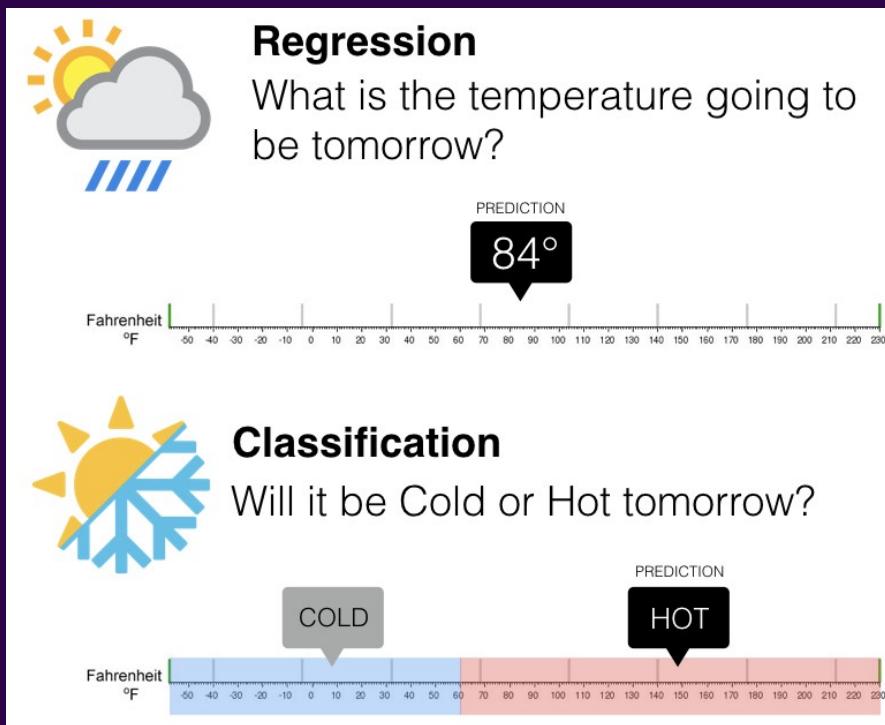
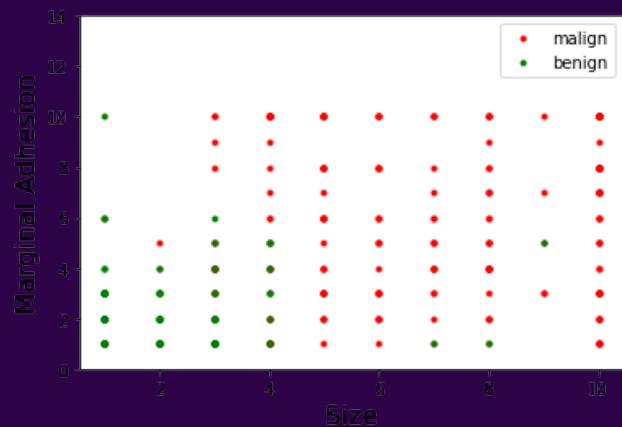
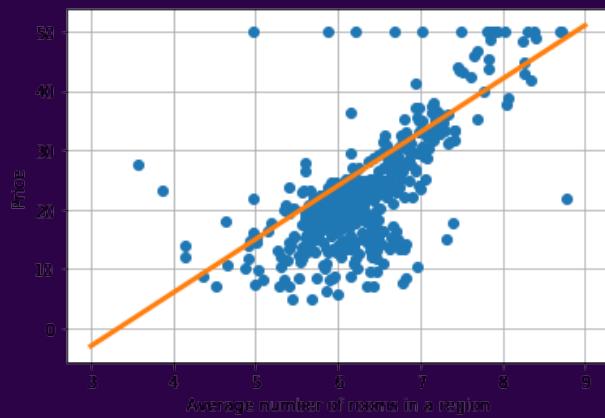


Figure: <https://www.pinterest.com/pin/672232681855858622/?lp=true> 

Classification Vs. Regression



(a) Breast cancer dataset

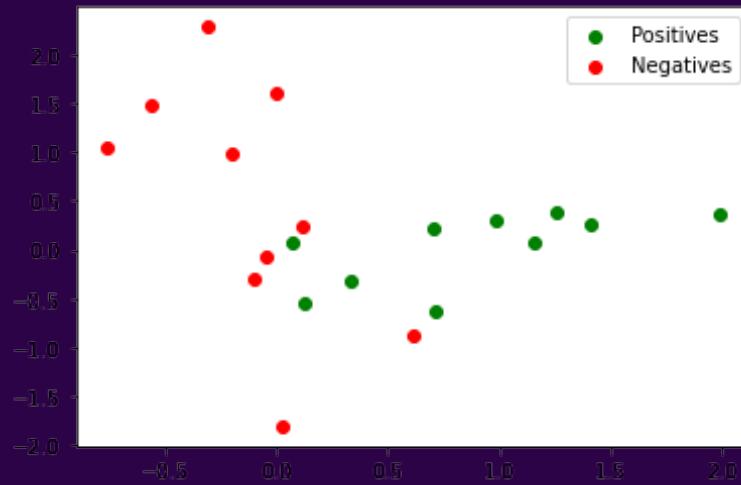


(b) Boston Housing dataset

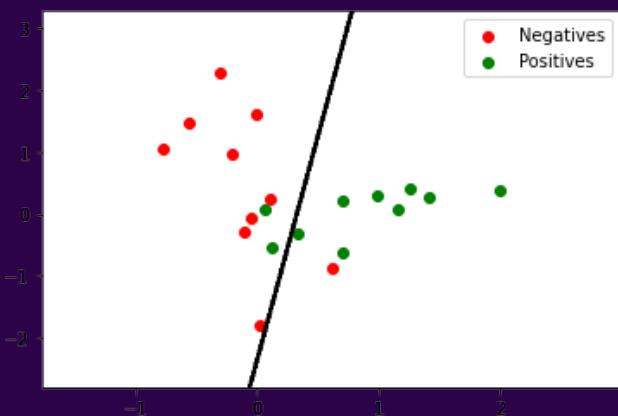
Classification

Given the dataset (x_i, y_i) for $i = 1, 2, \dots, N$, find a function $f(x)$ (model) so that it can predict the label \hat{y} for some input x , even if it is not in the dataset, i.e. $\hat{y} = f(x)$.

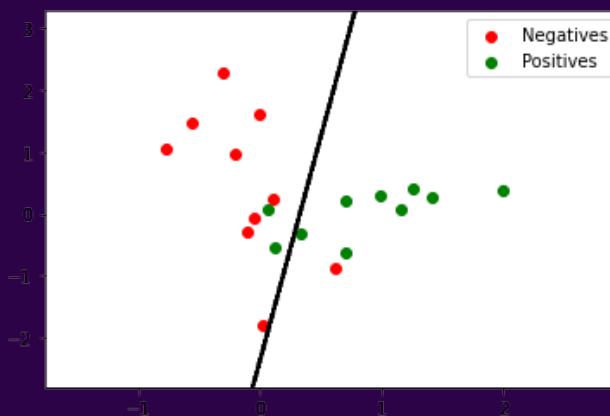
- Positive : $y = 1$
 - Negative : $y = 0$



Decision Boundary



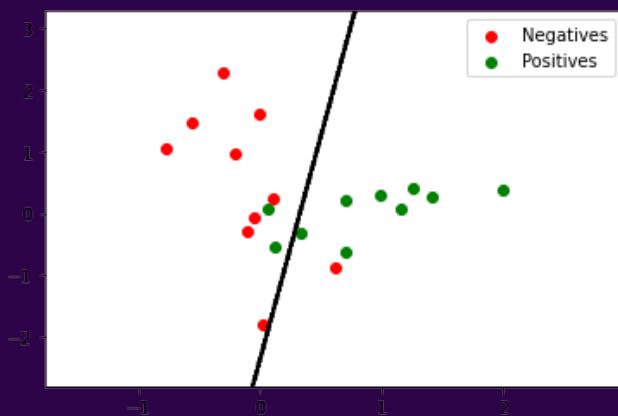
Decision Boundary



- Evaluation metric :

$$\text{Accuracy} = \frac{\text{Number of correct prediction}}{\text{Total number of prediction}}$$

- What is the accuracy in this example ?



■ Evaluation metric :

$$\text{Accuracy} = \frac{\text{Number of correct prediction}}{\text{Total number of prediction}} = \frac{17}{20} = 0.85 = 85\%$$

Need for a new model

- What would happen if we used the linear regression model :

$$\hat{y} = w_0 + w_1 x$$

Need for a new model

- What would happen if we used the linear regression model :

$$\hat{y} = w_0 + w_1 x$$

- y is 0 or 1
- \hat{y} will take any value between $-\infty$ and ∞

Need for a new model

- What would happen if we used the linear regression model :

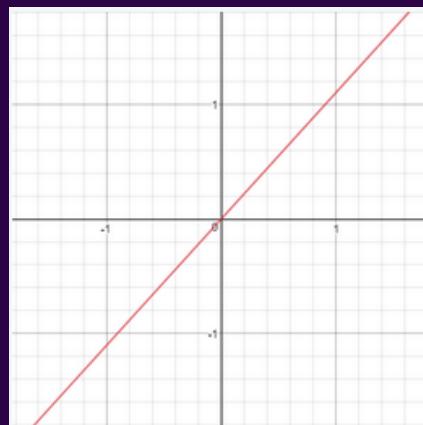
$$\hat{y} = w_0 + w_1 x$$

- y is 0 or 1
- \hat{y} will take any value between $-\infty$ and ∞
- It will be hard to find w_0 and w_1 that make the prediction \hat{y} match the label y .

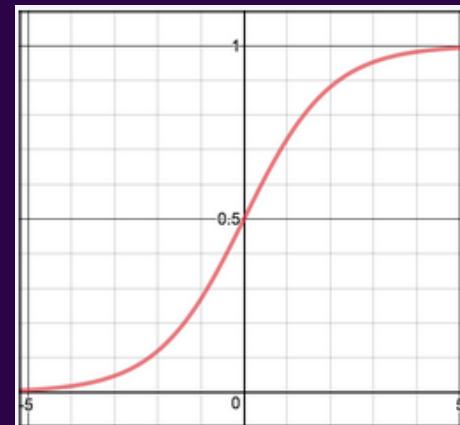
Sigmoid Function

- By applying the sigmoid function, we enforce $0 \leq \hat{y} \leq 1$

$$\hat{y} = \text{sigmoid}(w_0 + w_1 x) = \frac{1}{1 + e^{-(w_0 + w_1 x)}}$$



(a) Linear model



(b) Sigmoid model

A new loss function

- Binary cross entropy loss :

$$\text{Loss} = \frac{1}{N} \sum_{i=1}^N \left[-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i) \right]$$

pause

- What happens if $y_i = 0$:

$$\left[-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i) \right] = ?$$

A new loss function

- Binary cross entropy loss :

$$\text{Loss} = \frac{1}{N} \sum_{i=1}^N \left[-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i) \right]$$

- If $y_i = 0$:

$$\left[-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i) \right] = -\log(1 - \hat{y}_i)$$

A new loss function

■ Binary cross entropy loss :

$$\text{Loss} = \frac{1}{N} \sum_{i=1}^N \left[-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i) \right]$$

■ If $y_i = 0$:

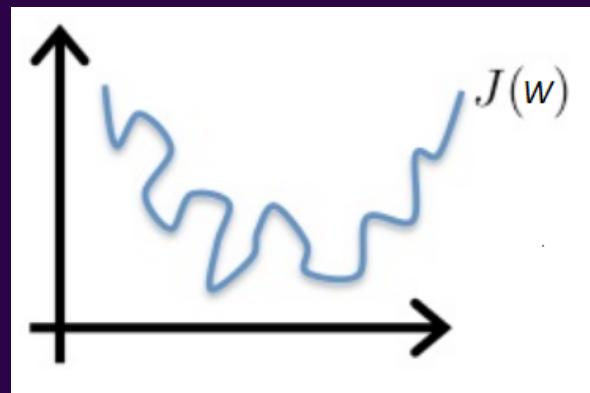
$$\left[-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i) \right] = -\log(1 - \hat{y}_i)$$

■ If $y_i = 1$:

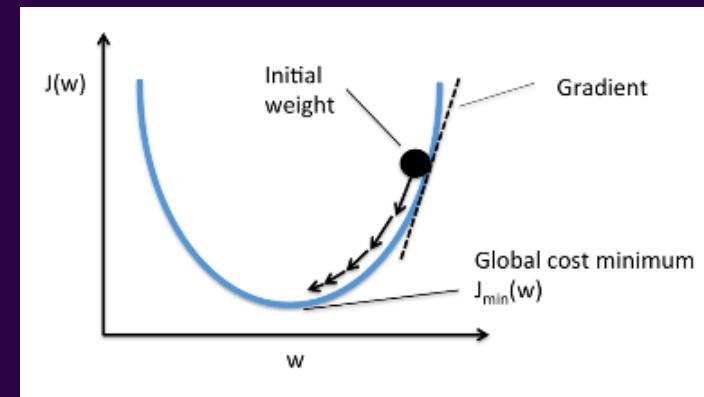
$$\left[-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i) \right] = -\log(\hat{y}_i)$$

MSE vs Binary cross entropy loss

- MSE of a logistic function has many local minima.
- The Binary cross entropy loss has only one minimum.



(a) MSE



(b) Binary cross entropy loss

Classifier

$$\hat{y} = \text{sigmoid}(w_0 + w_1 x) = \frac{1}{1 + e^{-(w_0 + w_1 x)}}$$

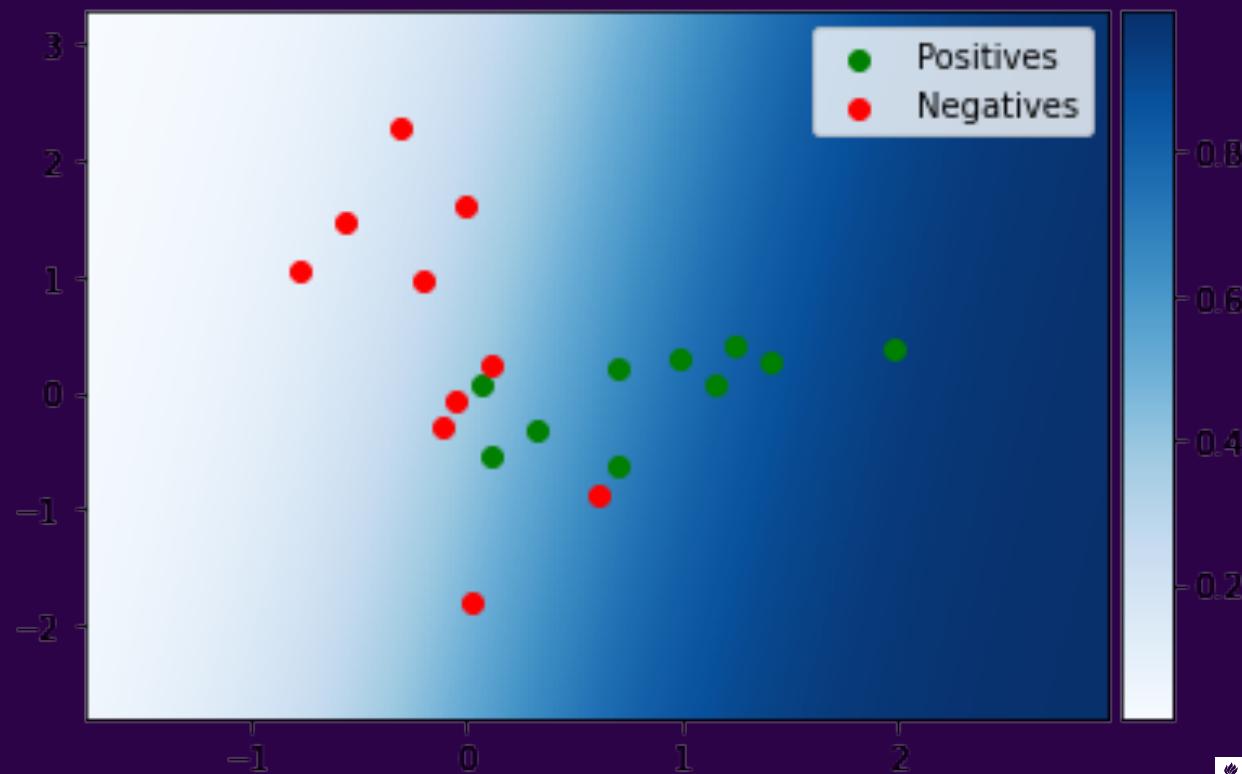
- How to deal with uncertainty ?
 - Thanks to the sigmoid, $\hat{y} = f(x)$ is between 0 and 1.

Classifier

$$\hat{y} = \text{sigmoid}(w_0 + w_1 x) = \frac{1}{1 + e^{-(w_0 + w_1 x)}}$$

- How to deal with uncertainty ?
 - Thanks to the sigmoid, $\hat{y} = f(x)$ is between 0 and 1.
- If \hat{y} is close to 0, the data is probably negative
- If \hat{y} is close to 1, the data is probably positive
- If \hat{y} is around 0.5, we are not sure.

Classifier



Decision Boundary

- Once, we have a classifier outputting a score $0 < \hat{y} < 1$, we need to create a decision rule.

Decision Boundary

- Once, we have a classifier outputting a score $0 < \hat{y} < 1$, we need to create a decision rule.
- Let $0 < t < 1$ be a Threshold :
 - If $\hat{y} > t$, \hat{y} is classified as positive.
 - If $\hat{y} < t$, \hat{y} is classified as negative.

Decision Boundary

- Once, we have a classifier outputting a score $0 < \hat{y} < 1$, we need to create a decision rule.
- Let $0 < t < 1$ be a Threshold :
 - If $\hat{y} > t$, \hat{y} is classified as positive.
 - If $\hat{y} < t$, \hat{y} is classified as negative.
- How to choose t ?

Impact of the threshold

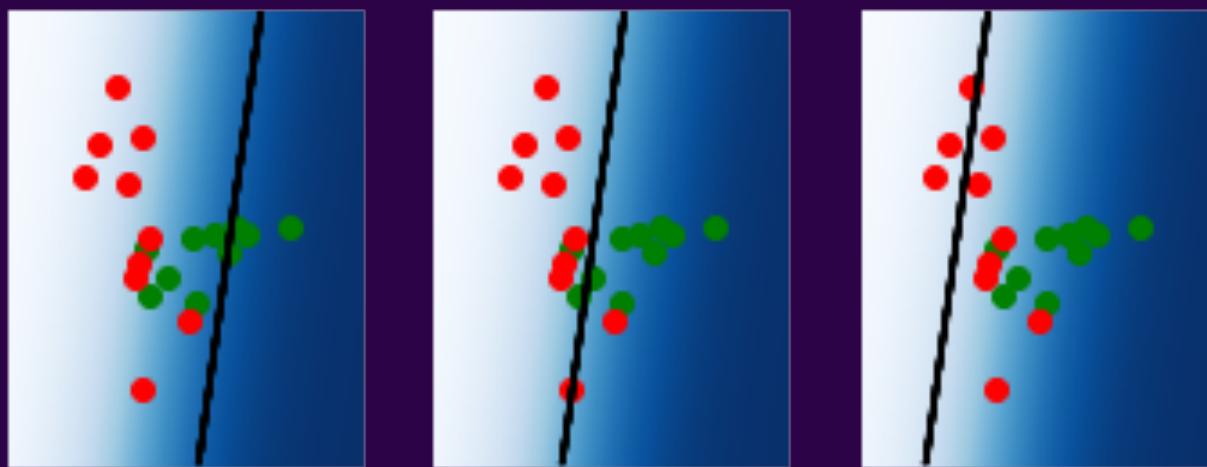


Figure: $t = 0.2, 0.5, 0.8$

Performance metrics for a classifier

- Accuracy of a classifier: percentage of correct classification
- Why accuracy alone is not a good measure for assessing the model ?

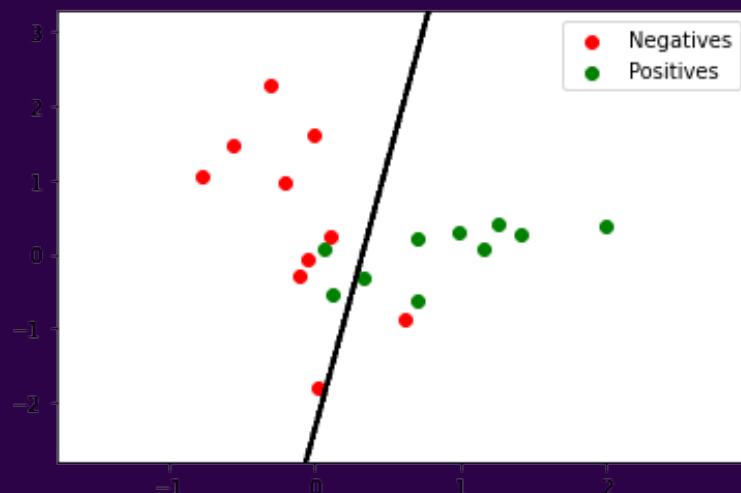
Performance metrics for a classifier

- Accuracy of a classifier: percentage of correct classification
- Why accuracy alone is not a good measure for assessing the model ?
 - Example: A rare disease occurs 1 in ten thousand people
 - A test that classifies everyone as free of the disease can achieve 99.999% accuracy when tested with people drawn randomly from the entire population

Types of Errors in Classification

- Correct predictions:
 - True Positive (TP) : Predict $\hat{y} = 1$ when $y = 1$
 - True Negative (TN) : Predict $\hat{y} = 0$ when $y = 0$
- Two types of errors:
 - False Positive/ False Alarm (FP): $\hat{y} = 1$ when $y = 0$
 - False Negative/ Missed Detection (FN): $\hat{y} = 0$ when $y = 1$

Example



- How many True Positive (TP) are there ?
- How many True Negative (TN) are there ?
- How many False Positive (FP) are there ?
- How many False Negative (FN) are there ?

Other metrics

- Sensitivity/Recall/TPR (How many positives are detected among all positive?)

$$\frac{\text{TP}}{\text{TP} + \text{FN}}$$

- Precision (How many detected positives are actually positive?)

$$\frac{\text{TP}}{\text{TP} + \text{FP}}$$

Review
o

Logistic
oooooooooooooooooooo

Demo
●○

Multiclass
oooo

Lab
oo

Outline

1 Review

2 Logistic Regression

3 Lab: Diagnosing Breast Cancer

4 Multiclass Classification

5 Lab: Iris Dataset

Lab: Diagnosing Breast Cancer

- We're going to use the breast cancer dataset to predict whether the patients' scans show a malignant tumour or a benign tumour.
- Let's try to find the best linear classifier using logistic regression.

Review
o

Logistic
oooooooooooooooooooo

Demo
oo

Multiclass
●ooo

Lab
oo

Outline

1 Review

2 Logistic Regression

3 Lab: Diagnosing Breast Cancer

4 Multiclass Classification

5 Lab: Iris Dataset

Multiclass Classification

- Previous model: $f(\mathbf{x}) = \sigma(\mathbf{w}^T \phi(\mathbf{x}))$
- Representing Multiple Classes:
 - One-hot / 1-of-K vectors, ex : 4 Class
 - Class 1 : $\mathbf{y} = [1, 0, 0, 0]$
 - Class 2 : $\mathbf{y} = [0, 1, 0, 0]$
 - Class 3 : $\mathbf{y} = [0, 0, 1, 0]$
 - Class 4 : $\mathbf{y} = [0, 0, 0, 1]$

Multiclass Classification

- Previous model: $f(\mathbf{x}) = \sigma(\mathbf{w}^T \phi(\mathbf{x}))$
 - Representing Multiple Classes:
 - One-hot / 1-of-K vectors, ex : 4 Class
 - Class 1 : $\mathbf{y} = [1, 0, 0, 0]$
 - Class 2 : $\mathbf{y} = [0, 1, 0, 0]$
 - Class 3 : $\mathbf{y} = [0, 0, 1, 0]$
 - Class 4 : $\mathbf{y} = [0, 0, 0, 1]$
 - Multiple outputs: $f(\mathbf{x}) = \text{softmax}(W^T \phi(\mathbf{x}))$
 - Shape of $W^T \phi(\mathbf{x})$: $(K, 1) = (K, D) \times (D, 1)$
 - $\text{softmax}(\mathbf{z})_k = \frac{e^{z_k}}{\sum_j e^{z_j}}$

Multiclass Classification

- Multiple outputs: $f(\mathbf{x}) = \text{softmax}(\mathbf{z})$ with $\mathbf{z} = W^T \phi(\mathbf{x})$

- $\text{softmax}(\mathbf{z})_k = \frac{e^{z_k}}{\sum_j e^{z_j}}$

- Softmax example: If $\mathbf{z} = \begin{bmatrix} -1 \\ 2 \\ 1 \\ -4 \end{bmatrix}$ then,

$$\text{softmax}(\mathbf{z}) = \begin{bmatrix} \frac{e^{-1}}{e^{-1} + e^2 + e^1 + e^{-4}} \\ \frac{e^2}{e^{-1} + e^2 + e^1 + e^{-4}} \\ \frac{e^1}{e^{-1} + e^2 + e^1 + e^{-4}} \\ \frac{e^{-4}}{e^{-1} + e^2 + e^1 + e^{-4}} \end{bmatrix} \approx \begin{bmatrix} 0.035 \\ 0.704 \\ 0.259 \\ 0.002 \end{bmatrix}$$

Cross-entropy

- Multiple outputs: $\hat{\mathbf{y}}_i = \text{softmax}(W^T \phi(\mathbf{x}_i))$
- Cross-Entropy: $J(W) = - \sum_{i=1}^N \sum_{k=1}^K \mathbf{y}_{ik} \log(\hat{\mathbf{y}}_{ik})$
- Example : $K = 4$

If, $\mathbf{y}_i = [0, 0, 1, 0]$ then, $\sum_{k=1}^K \mathbf{y}_{ik} \log(\hat{\mathbf{y}}_{ik}) = \log(\hat{\mathbf{y}}_{i3})$

Review
○

Logistic
oooooooooooooooooooo

Demo
oo

Multiclass
oooo

Lab
●○

Outline

1 Review

2 Logistic Regression

3 Lab: Diagnosing Breast Cancer

4 Multiclass Classification

5 Lab: Iris Dataset

Review
o

Logistic
oooooooooooooooooooo

Demo
oo

Multiclass
oooo

Lab
oo●

Lab: Iris Dataset

- Open demo_iris.ipynb