MACHINE LEARNING

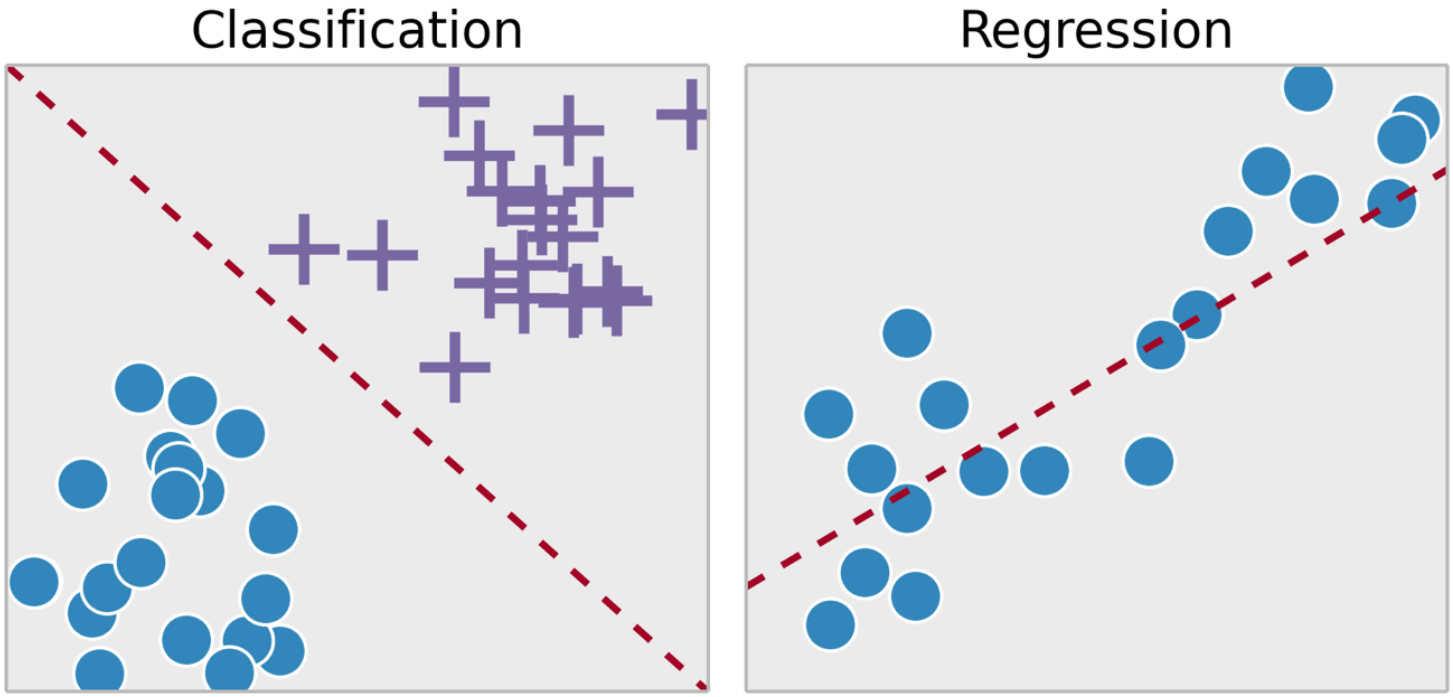
Lecture 4

NUACA

2017

SUPERVISED LEARNING

In supervised learning, the training data consists of input and output.



CLASSIFICATION

Email: Spam / Not Spam

Online Transaction: Fraudulent (Yes/No)

Tumor: Malignant/ Benign

$y \in \{0, 1\}$ – binary classification

$y \in \{0, 1, 2, 3, \ldots, n\}$ – multiclass classification

LOGISTIC REGRESSION

Logistic Regression is **not** a method for regression. It is for (binary) classification!

It allows choosing the target class by computing the probability of each category.

LOGISTIC REGRESSION

Recall Linear Regression:

$$h(\mathbf{x}) = w_o x_0 + w_1 x_1 + \cdots + w_D x_D = \sum_{j=1}^{D} w_j x_j = \mathbf{w^T x}$$

where

$$\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \cdots \\ x_{D+1} \end{bmatrix} \in \mathbb{R}^{D+1}, \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \cdots \\ w_{D+1} \end{bmatrix} \in \mathbb{R}^{D+1}$$

$h(\mathbf{x}_i)$ can be $(-\infty, \infty)$.

But probabilities have to be within [0, 1] range

LOGISTIC REGRESSION

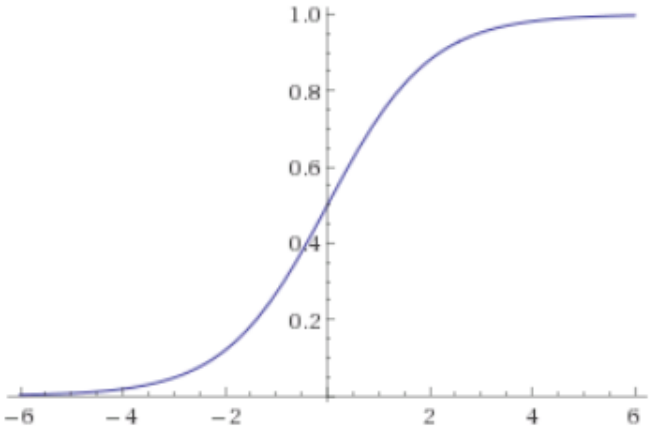
We want $0 \leq h(\mathbf{x}) \leq 1$.

We can achieve this by doing the following

$$h(\mathbf{x}) = \sigma(\mathbf{w}^{\mathbf{T}}\mathbf{x}),$$

where σ is called the sigmoid or logistic function:

$$\sigma(z) = \frac{1}{1+\exp^{-z}}$$

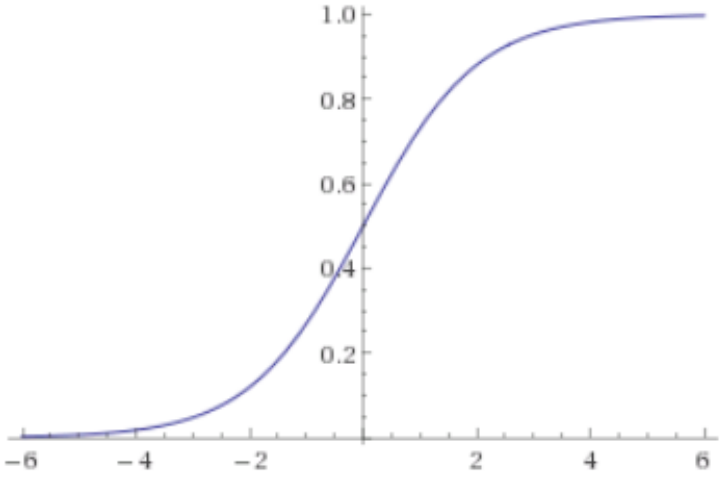


As $z \to -\infty, \sigma(z) \to 0,$

$z \to +\infty, \sigma(z) \to 1$

It is define on $(-\infty, \infty)$ and has non-negative derivative everywhere.

LOGISTIC REGRESSION



Thus, we'll have

$$h(\mathbf{x}) = \frac{\mathbf{1}}{\mathbf{1}+e^{-\mathbf{w}^T\mathbf{x}}}$$

$h(\mathbf{x})$ will thus be within the [0, 1] range.

Therefore, we can be used to estimate the probability that y=1 on input **x**

Example: $\mathbf{x} = [x_0 \;\; x_1]^T = [1 \;\; tumorSize]^T$

if $h(\mathbf{x}) = \mathbf{0.7}$, tell patient that 70% chance of tumor being malignant

PROBABILITY THEORY NOTATIONS

Formally, the probability of the input **x** belonging to the class 1 is:

$$\mathbb{P}[y = 1|\mathbf{x}; \mathbf{w}] = \frac{1}{1 + \exp^{-\mathbf{w}^T\mathbf{x}}} = \sigma(\mathbf{w}^T\mathbf{x})$$

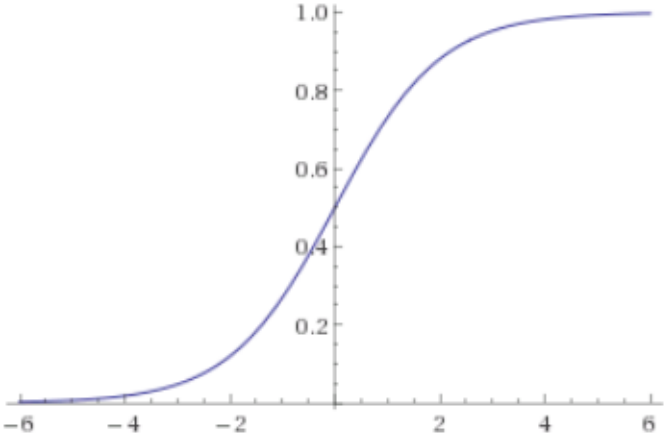
the probability of the input **x** belonging to the class 0 is:

$$\mathbb{P}[y = 0|\mathbf{x}; \mathbf{w}] = 1 - \mathbb{P}[y = 1|\mathbf{x}; \mathbf{w}] = 1 - \sigma(\mathbf{w}^T\mathbf{x})$$

The formulas above can be written as a single formula:

$$\mathbb{P}[y; \mathbf{x}; \mathbf{w}] = \sigma(\mathbf{w}^T\mathbf{x})^y(1 - \sigma(\mathbf{w}^T\mathbf{x}))^{1-y}$$

PREDICTIONS



$$\mathbb{P}[y = 1|\mathbf{x}; \mathbf{w}] = \frac{1}{1 + \exp^{-\mathbf{w}^T\mathbf{x}}} = \sigma(\mathbf{w}^T\mathbf{x})$$

We will predict $y = 1$ if $\mathrm{P}[\mathrm{y} = 1; \mathbf{x}; \mathbf{w}] \geq 0.5$

This means $\sigma(\mathbf{w}^\mathbf{T}\mathbf{x}) \geq 0.5$, which also means $\mathbf{w}^\mathbf{T}\mathbf{x} \geq 0$

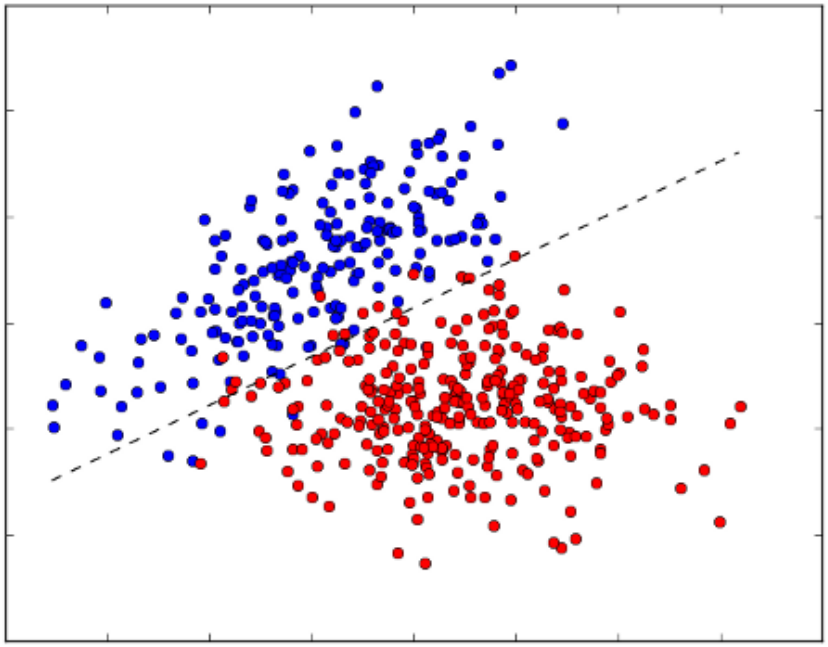
We will predict $y = 0$ if $\mathrm{P}[\mathrm{y} = 1; \mathbf{x}; \mathbf{w}] < 0.5$

This means $\left(\mathbf{w}^\mathbf{T}\mathbf{x}\right) < 0.5$, which also means $\mathbf{w}^\mathbf{T}\mathbf{x} < 0$

GEOMETRIC MEANING

Thus, the sign of $\mathbf{w}^{\mathbf{T}}\mathbf{x}$ determines whether the target is 0 or 1.

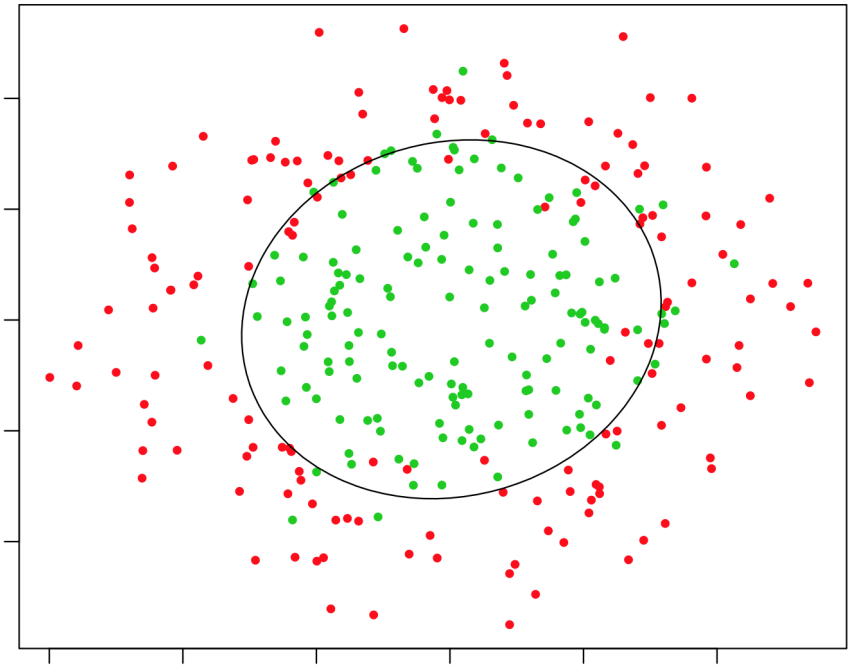
As $\mathbf{w}^{\mathbf{T}}\mathbf{x} = w_o x_0 + w_1 x_1 + w_2 x_2$ is a line, logistic regression will produce a linear decision boundary.



POLYNOMIAL FEATURES

If we use polynomial features instead, get more complex decision boundries.

$$\mathbf{w}^{\mathbf{T}}\mathbf{x} = w_o x_0 + w_1 x_1 + w_2 x_2 + w_3 {x_1}^2 + w_4 {x_2}^2 + w_5 x_1 x_2$$




TIME FOR A BREAK

COST/LOSS FUNCTION

Recall the Cost Function of Linear Regression:

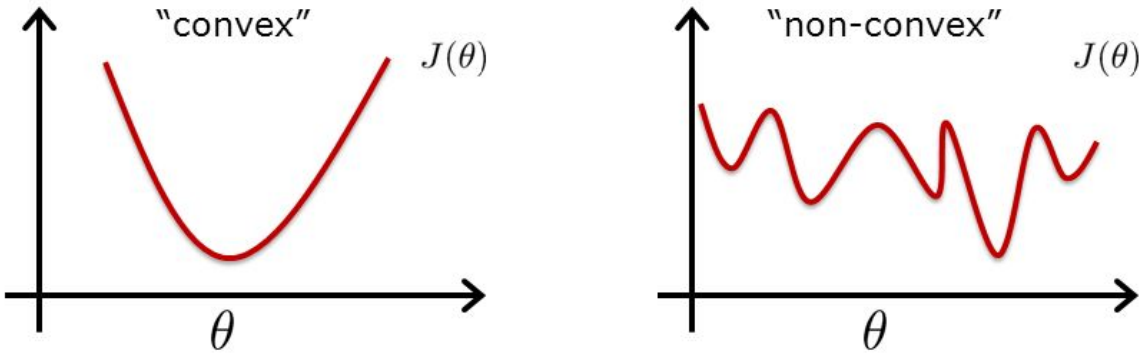
$$\mathcal{L}(\boldsymbol{w}) = \mathcal{L}(w_o, w_1) = \frac{1}{2N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

Here, if for predictions we use the following formula

$$\hat{y} = \mathbb{I}(\sigma(\mathbf{w}^T\mathbf{x}) \geq 0.5) = \mathbb{I}(\mathbf{w}^T\mathbf{x} \geq 0)$$

we'll end up with a non-convex function:

Trust me, you don't want non-convex optimisation



COST/LOSS FUNCTION

Let's define a Loss function for each prediction:

$$Cost(h(\mathbf{x}), y) = \begin{cases} -\log\big(h(\mathbf{x})\big) & if \;\; y = 1 \\ -\log\big(1 - h(\mathbf{x})\big) & if \;\; y = 0 \end{cases}$$

I bet you're wondering, why on earth this makes sense.

Let's clarify on the board!!

The Cost function can be written in a single formula:

$$Cost(h(\mathbf{x}), y) = -\,\mathrm{y}\,\log h(\mathbf{x}) - (1 - y)\log\big(1 - h(\mathbf{x})\big)$$

- Hint: consider y=0, y =1 cases

COST FUNCTION

Our goal now would be to minimise the following objective:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N} Cost(h(\mathbf{x_i}), y)$$

$$= \frac{1}{N}\sum_{i=1}^{N}(-\mathrm{y}\log h(\mathbf{x_i}) - (1-y)\log\big(1-h(\mathbf{x_i})\big))$$

$$= -\frac{1}{N}\sum_{i=1}^{N}(\mathrm{y}\log\sigma\big(\mathbf{w^T x_i}\big) + (1-y)\log\big(1-\sigma(\mathbf{w^T x_i})\big))$$

which is convex! (Try to prove it)

GRADIENT DESCENT

We can now use Gradient Descent algorithm to find the optimal parameters: $\mathbf{w}$

Repeat for i = 0,1,…, *D* until convergence:

$$w_i^{(t+1)} := w_i^{(t)} - \alpha \frac{\partial}{\partial w_i} \mathcal{L}(\mathbf{w})$$

Update all at the same time. α is called the learning rate.

In the vector form, Gradient Descent will look like this:

$$\mathbf{w}^{(t+1)} := \mathbf{w}^{(t)} - \alpha \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w})$$

GRADIENT DESCENT

Repeat for i = 0,1,…, *D* until convergence:

$$w_i^{(t+1)} := w_i^{(t)} - \alpha \frac{\partial}{\partial w_i} \mathcal{L}(\mathbf{w}) \quad (1)$$

Where $\mathcal{L}(\mathbf{w}) = -\frac{1}{N} \sum_{i=1}^{N} (y \log \sigma(\mathbf{w}^T \mathbf{x_i}) + (1-y) \log(1 - \sigma(\mathbf{w}^T \mathbf{x_i}))$ (2)

Combining equations (1) and (2), we'll get

Repeat for i = 0,1,…, *D* until convergence:

$$w_i^{(t+1)} := w_i^{(t)} - \alpha \sum_{i=1}^{N} (\sigma(\mathbf{w}^T \mathbf{x_i}) - y)\, x_{ij}$$

$$= w_i^{(t)} - \alpha \sum_{i=1}^{N} (\hat{y} - y)\, x_{ij}$$

Which looks awful like the gradient update of Linear Regression.

LITERATURE

- (Murphy) Chapter 8.1-3, 8.6
- https://en.wikipedia.org/wiki/Logistic_regression

QUESTIONS?