Classification I

COMP9417, 23T2

Section 1

Classification

Classification

Recall the standard form of a machine learning problem:

- We have 'input' data X and targets/outputs y
- Our data can be modelled as $y = f(X)$
- Goal is to find the best approximation for f as $\hat{f}$

Here, $f(x)$ outputs *classes* rather than numeric values.

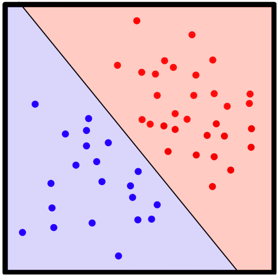
Note:

We call a two-class problem a binary classification problem.

Linearly Separable Datasets

We define a **linearly separable** dataset as one which can be classified in a binary fashion using a hyperplane.

More simply, if you can classify it by drawing a line through it. Your dataset is linearly separable.



Section 2

Perceptron

Perceptron

Learns weights w for a decision boundary $w^T\mathbf{x} = 0$, where $\mathbf{x}$ represents points on the Cartesian plane, not our dataset.

Key Properties

- The classic perceptron solves only binary classification
- **Always converges** to a solution if the dataset in linearly separable
- Solutions can differ depending on starting weights and learning rate

The algorithm

For weights w and a learning rate η.

$converged \leftarrow 0$
while not $converged$ **do**
  $converged \leftarrow 1$
  for $x_i \in X, y_i \in y$ **do**
    if $y_i w \cdot x_i \leq 0$ **then**
      $w \leftarrow w + \eta y_i x_i$
      $converged \leftarrow 0$
    end if
  end for
end while

Section 3

Logistic Regression

Logistic Regression

Often called *logit* **model**. A way for us to use a linear combination $w^T x$ to predict probabilities of a binary classification problem.

For a data point (x_i, y_i) the model will predict:

$$P(y_i = 1 | x_i)$$

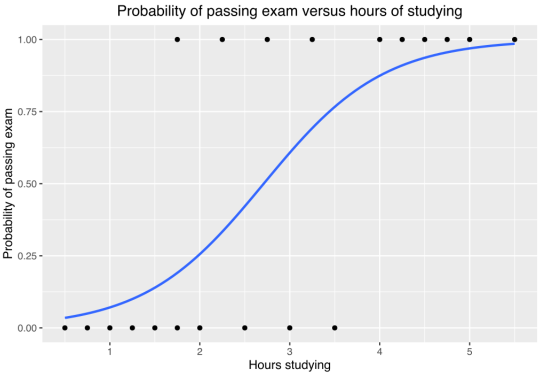
Simply, the probability that the target belongs to class 1 given the datapoint at index i.

The logistic regression is defined as the following function:

$$\sigma(w^T x_i) = \frac{1}{1 + e^{-w^T x_i}}$$

In the basic case where we only have one feature:

$$\sigma(w^T x_i) = \frac{1}{1 + e^{-w_0 - w_1 x_i}}$$



Section 4

3 (a, b, c)

3 (a, b, c)

If we define the binary prediction problem as a probability:

$$P(y = 1|x) = p(x)$$

We write the logistic regression prediction as:

$$\hat{p}(x) = \sigma(\hat{w}^T x)$$

$$\text{where } \sigma(z) = \frac{1}{1 + e^{-z}}$$

where we predict the class of an input x to be 1 if $\hat{p}(x) \geq 0.5$.

3a

What is the role of the sigmoid function here?

3a

What is the role of the sigmoid function here?

In a linear model, we can't simply predict probabilities or classes with the classic equation $\hat{p}(x) = \hat{w}^T x$.

3a

What is the role of the sigmoid function here?

In a linear model, we can't simply predict probabilities or classes with the classic equation $\hat{p}(x) = \hat{w}^T x$. The sigmoid $\sigma(z)$ us model probabilities in a valid interval ($[0, 1]$).

3b

Consider the statistical view of the binary classification problem $y_i|x_i \sim \text{Bernoulli}(p_i^*)$ where $p_i^* = \sigma(x_i^T w)$ is our logistic regression model.

By definition of the Bernoulli:

$$P(y|x) = p^y(1-p)^{1-y}$$

So, we can estimate p using MLE:

3b

Consider the statistical view of the binary classification problem $y_i | x_i \sim \text{Bernoulli}(p_i^*)$ where $p_i^* = \sigma(x_i^T w)$ is our logistic regression model.

By definition of the Bernoulli:

$$P(y|x) = p^y (1-p)^{1-y}$$

So, we can estimate p using MLE:

$$\ln L(w) = \ln \left(\prod_{i=1}^{n} P(y_i | x_i) \right)$$

3b

Consider the statistical view of the binary classification problem $y_i|x_i \sim \text{Bernoulli}(p_i^*)$ where $p_i^* = \sigma(x_i^T w)$ is our logistic regression model.

By definition of the Bernoulli:

$$P(y|x) = p^y(1-p)^{1-y}$$

So, we can estimate p using MLE:

$$\ln L(w) = \ln\left(\prod_{i=1}^{n} P(y_i|x_i)\right)$$

$$= \sum_{i=1}^{n} \ln P(y_i|x_i)$$

$$= \sum_{i=1}^{n} \ln \left(p_i^{y_i} (1 - p_i)^{1-y_i} \right)$$

$$= \sum_{i=1}^{n} \ln \left(p_i^{y_i} (1 - p_i)^{1-y_i} \right)$$

$$= \sum_{i=1}^{n} [y_i \ln p_i + (1 - y_i) \ln(1 - p_i)]$$

$$= \sum_{i=1}^{n} \ln\left(p_i^{y_i}(1-p_i)^{1-y_i}\right)$$

$$= \sum_{i=1}^{n} [y_i \ln p_i + (1-y_i)\ln(1-p_i)]$$

$$= \sum_{i=1}^{n} \left[y_i \ln\left(\sigma(w^T x_i)\right) + (1-y_i)\ln\Big(1-\sigma(w^T x_i)\Big)\right]$$

$$= \sum_{i=1}^{n} \ln\left(p_i^{y_i}(1-p_i)^{1-y_i}\right)$$

$$= \sum_{i=1}^{n} [y_i \ln p_i + (1-y_i)\ln(1-p_i)]$$

$$= \sum_{i=1}^{n} \left[y_i \ln\left(\sigma(w^T x_i)\right) + (1-y_i)\ln\Big(1-\sigma(w^T x_i)\Big)\right]$$

$$= \sum_{i=1}^{n} \left[y_i \ln\left(\frac{\sigma(w^T x_i)}{1-\sigma(w^T x_i)}\right) + \ln\Big(1-\sigma(w^T x_i)\Big)\right]$$

We then solve,

$$\hat{w}_{\mathsf{MLE}} = \arg\max_w \sum_{i=1}^{n} \left[y_i \ln\left(\frac{\sigma(w^T x_i)}{1 - \sigma(w^T x_i)} \right) + \ln\Big(1 - \sigma(w^T x_i)\Big) \right]$$

Section 5

4 (a, b, c)

4 (a, b, c)

We have the following optimisation problem,

$$\begin{aligned}\hat{w} &= \arg\min_{w} L(w) \\ &= \arg\min_{w} -\sum_{i=1}^{n} \left[y_i \ln\left(p_i\right) + \left(1 - y_i\right) \ln(1 - p_i)\right]\end{aligned}$$

where $p_i = \sigma(w^T x_i)$.

4a

Show that $\sigma(z)' = \sigma(z)(1 - \sigma(z))$ **where** $\sigma(z) = \frac{1}{1+e^{-z}}$

4a

Show that $\sigma(z)' = \sigma(z)(1 - \sigma(z))$ **where** $\sigma(z) = \frac{1}{1+e^{-z}}$

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

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

4a

Show that $\sigma(z)' = \sigma(z)(1 - \sigma(z))$ **where** $\sigma(z) = \frac{1}{1+e^{-z}}$

$$\begin{aligned}
\sigma(z) &= \frac{1}{1 + e^{-z}} \\
\sigma'(z) &= \frac{-e^{-z}}{(1 + e^{-z})^2} \\
&= \frac{-e^{-z} - 1 + 1}{(1 + e^{-z})^2}
\end{aligned}$$

4a

Show that $\sigma(z)' = \sigma(z)(1 - \sigma(z))$ **where** $\sigma(z) = \frac{1}{1+e^{-z}}$

$$
\begin{aligned}
\sigma(z) &= \frac{1}{1+e^{-z}} \\
\sigma'(z) &= \frac{-e^{-z}}{(1+e^{-z})^2} \\
&= \frac{-e^{-z}-1+1}{(1+e^{-z})^2} \\
&= \frac{1}{(1+e^{-z})^2} - \frac{1}{1+e^{-z}} \\
&= \sigma(z)(1-\sigma(z))
\end{aligned}
$$

Therefore, show that $\frac{dp_i}{dw} = p_i(1 - p_i)w_i$

Therefore, show that $\frac{dp_i}{dw} = p_i(1 - p_i)w_i$

$$\begin{aligned}
\frac{dp_i}{dw} &= \frac{dp_i}{d(w^T x_i)} \frac{d(w^T x_i)}{dw} \\
&= p_i(1 - p_i) \frac{d(w^T x_i)}{dw} \\
&= p_i(1 - p_i)x_i
\end{aligned}$$

4b

Use the previous result to show that $\frac{d \ln p_i}{dw} = (1 - p_i)x_i$

4b

Use the previous result to show that $\frac{d \ln p_i}{dw} = (1 - p_i)x_i$

$$\begin{aligned}
\frac{d \ln p_i}{dw} &= \frac{d \ln p_i}{dp_i}\frac{dp_i}{dw} \\
&= \frac{1}{p_i} p_i (1 - p_i) x_i \\
&= (1 - p_i) x_i
\end{aligned}$$

4c

Using the previous results compute the gradient $\frac{dL(w)}{dw}$

$$\frac{dL(w)}{dw} = \frac{d}{dw}\left[-\sum_{i=1}^{n} y_i \ln p_i + (1 - y_i)\ln(1 - p_i)\right]$$

4c

Using the previous results compute the gradient $\frac{dL(w)}{dw}$

$$\frac{dL(w)}{dw} = \frac{d}{dw}\left[-\sum_{i=1}^{n} y_i \ln p_i + (1-y_i)\ln(1-p_i)\right]$$

$$= -\sum_{i=1}^{n}\left[y_i \frac{d}{dw}\ln p_i + (1-y_i)\frac{d}{dw}\ln(1-p_i)\right]$$

4c

Using the previous results compute the gradient $\frac{dL(w)}{dw}$

$$\frac{dL(w)}{dw} = \frac{d}{dw}\left[-\sum_{i=1}^{n} y_i \ln p_i + (1-y_i)\ln(1-p_i)\right]$$

$$= -\sum_{i=1}^{n}\left[y_i \frac{d}{dw}\ln p_i + (1-y_i)\frac{d}{dw}\ln(1-p_i)\right]$$

$$= -\sum_{i=1}^{n}\left[y_i(1-p_i)x_i + (1-y_i)\left(-\frac{1}{1-p_i}p_i(1-p_i)x_i\right)\right]$$

4c

Using the previous results compute the gradient $\frac{dL(w)}{dw}$

$$
\begin{aligned}
\frac{dL(w)}{dw} &= \frac{d}{dw}\left[-\sum_{i=1}^{n} y_i \ln p_i + (1-y_i)\ln(1-p_i)\right] \\
&= -\sum_{i=1}^{n}\left[y_i \frac{d}{dw}\ln p_i + (1-y_i)\frac{d}{dw}\ln(1-p_i)\right] \\
&= -\sum_{i=1}^{n}\left[y_i(1-p_i)x_i + (1-y_i)\left(-\frac{1}{1-p_i}p_i(1-p_i)x_i\right)\right] \\
&= -\sum_{i=1}^{n}(y_i(1-p_i)x_i - (1-y_i)p_i x_i)
\end{aligned}
$$

4c

Using the previous results compute the gradient $\frac{dL(w)}{dw}$

$$
\begin{aligned}
\frac{dL(w)}{dw} &= \frac{d}{dw}\left[-\sum_{i=1}^{n} y_i \ln p_i + (1-y_i)\ln(1-p_i)\right] \\
&= -\sum_{i=1}^{n}\left[y_i \frac{d}{dw}\ln p_i + (1-y_i)\frac{d}{dw}\ln(1-p_i)\right] \\
&= -\sum_{i=1}^{n}\left[y_i(1-p_i)x_i + (1-y_i)\left(-\frac{1}{1-p_i}p_i(1-p_i)x_i\right)\right] \\
&= -\sum_{i=1}^{n}(y_i(1-p_i)x_i - (1-y_i)p_i x_i) \\
&= -\sum_{i=1}^{n}(y_i - p_i)x_i
\end{aligned}
$$