# Machine Learning

## Discriminative Classification & Logistic Regression

Dariush Hosseini

dariush.hosseini@ucl.ac.uk
Department of Computer Science
University College London

# Lecture Overview

# Lecture Overview

By the end of this lecture you should:

1. Know the problem of **Classification** and understand the different paradigmatic approaches to its solution
2. Understand **logistic regression**, its motivation, and its use to solve the **binary classification** problem
3. Know that traditional logistic regression can be extended in a number of ways - both in terms of setting (the **Bayesian approach**) and in terms of application (**multinomial classification**)

# Lecture Overview

# Setting

- Recall that in classification problems we seek to **learn** a mapping between input features and a discrete output label
- We can then use this mapping to make output **predictions** given novel input data
- In **binary** classification the output set comprises 2 classes, while in **multinomial** classification the output comprises greater than 2 unordered (categorical) labels

# Notation

- **Inputs**
  $\mathbf{x} = [1, x_1, ..., x_m]^T \in \mathbb{R}^{m+1}$

- **Binary Outputs**
  $y \in \{0, 1\}$

- **Training Data**
  $\mathcal{S} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$

# Probabilistic Environment

- We assume:
  - $\mathbf{x}$ is the outcome of a random variable $\mathcal{X}$
  - $y$ is the outcome of a random variable $\mathcal{Y}$
  - $(\mathbf{x}, y)$ are drawn i.i.d. from some data generating distribution, $\mathcal{D}$, i.e.:

$$(\mathbf{x}, y) \sim \mathcal{D}$$

    and:

$$\mathcal{S} \sim \mathcal{D}^n$$

# Learning Problem

- **Representation**

$$f \in \mathcal{F}$$

- **Evaluation**
  - **Loss Measure:**

$$\mathcal{E}(f(\mathbf{x}), y) = \mathbb{I}[y \neq f(\mathbf{x})]$$

  - **Generalisation Loss:**

$$\mathsf{L}(\mathcal{E}, \mathcal{D}, f) = \mathbb{E}_{\mathcal{D}}\left[\mathbb{I}[\mathcal{Y} \neq f(\mathcal{X})]\right]$$

Where $\mathcal{D}$ is characterised by $p_{\mathcal{X},\mathcal{Y}}(\mathbf{x}, y) = p_{\mathcal{Y}}(y|\mathbf{x})p_{\mathcal{X}}(\mathbf{x})$ for some pmf, $p_{\mathcal{Y}}(\cdot|\cdot)$, and some pdf, $p_{\mathcal{X}}(\cdot)$

- **Optimisation**

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}\left[\mathbb{I}[\mathcal{Y} \neq f(\mathcal{X})]\right]$$

# Optimisation Problem

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}} \left[ \mathbb{I}[\mathcal{Y} \neq f(\mathcal{X})] \right]$$

# Optimisation Problem

$$
\begin{aligned}
f^* &= \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}\left[\mathbb{I}[\mathcal{Y} \neq f(\mathcal{X})]\right] \\
&= \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{y \in \{0,1\}} \int p_{\mathcal{Y}}(y|\mathbf{x}) p_{\mathcal{X}}(\mathbf{x}) \mathbb{I}[y \neq f(\mathbf{x})] d\mathbf{x}
\end{aligned}
$$

# Optimisation Problem

$$
\begin{aligned}
f^* &= \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}\left[\mathbb{I}[\mathcal{Y} \neq f(\mathcal{X})]\right] \\
&= \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{y \in \{0,1\}} \int p_{\mathcal{Y}}(y|\mathbf{x}) p_{\mathcal{X}}(\mathbf{x}) \mathbb{I}[y \neq f(\mathbf{x})] d\mathbf{x} \\
&= \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{y \in \{0,1\}} \int p_{\mathcal{Y}}(y|\mathbf{x}) p_{\mathcal{X}}(\mathbf{x}) \Big( f(\mathbf{x})(1-y) + (1-f(\mathbf{x}))y \Big) d\mathbf{x}
\end{aligned}
$$

# Optimisation Problem

$$
\begin{aligned}
f^* &= \underset{f \in \mathcal{F}}{\operatorname{argmin}} \, \mathbb{E}_{\mathcal{D}}\left[\mathbb{I}[\mathcal{Y} \neq f(\mathcal{X})]\right] \\
&= \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{y \in \{0,1\}} \int p_{\mathcal{Y}}(y|\mathbf{x}) p_{\mathcal{X}}(\mathbf{x}) \mathbb{I}[y \neq f(\mathbf{x})] d\mathbf{x} \\
&= \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{y \in \{0,1\}} \int p_{\mathcal{Y}}(y|\mathbf{x}) p_{\mathcal{X}}(\mathbf{x}) \Big( f(\mathbf{x})(1-y) + (1-f(\mathbf{x}))y \Big) d\mathbf{x} \\
&= \underset{f \in \mathcal{F}}{\operatorname{argmin}} \int \Big( p_{\mathcal{Y}}(y=1|\mathbf{x})(1-f(\mathbf{x})) + p_{\mathcal{Y}}(y=0|\mathbf{x}) f(\mathbf{x}) \Big) p_{\mathcal{X}}(\mathbf{x}) d\mathbf{x}
\end{aligned}
$$

# Optimisation Problem

$$
\begin{aligned}
f^* &= \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}\left[\mathbb{I}[\mathcal{Y} \neq f(\mathcal{X})]\right] \\
&= \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{y \in \{0,1\}} \int p_{\mathcal{Y}}(y|\mathbf{x}) p_{\mathcal{X}}(\mathbf{x}) \mathbb{I}[y \neq f(\mathbf{x})] d\mathbf{x} \\
&= \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{y \in \{0,1\}} \int p_{\mathcal{Y}}(y|\mathbf{x}) p_{\mathcal{X}}(\mathbf{x}) \Big( f(\mathbf{x})(1-y) + (1 - f(\mathbf{x}))y \Big) d\mathbf{x} \\
&= \operatorname*{argmin}_{f \in \mathcal{F}} \int \Big( p_{\mathcal{Y}}(y=1|\mathbf{x})(1 - f(\mathbf{x})) + p_{\mathcal{Y}}(y=0|\mathbf{x}) f(\mathbf{x}) \Big) p_{\mathcal{X}}(\mathbf{x}) d\mathbf{x} \\
&= \operatorname*{argmin}_{f \in \mathcal{F}} \int \Big( p_{\mathcal{Y}}(y=1|\mathbf{x})(1 - f(\mathbf{x})) + (1 - p_{\mathcal{Y}}(y=1|\mathbf{x})) f(\mathbf{x}) \Big) p_{\mathcal{X}}(\mathbf{x}) d\mathbf{x}
\end{aligned}
$$

# Optimisation Problem

$$
\begin{aligned}
f^* &= \operatorname*{argmin}_{f\in\mathcal{F}} \mathbb{E}_{\mathcal{D}}\left[\mathbb{I}[\mathcal{Y} \neq f(\mathcal{X})]\right] \\
&= \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{y\in\{0,1\}} \int p_{\mathcal{Y}}(y|\mathbf{x}) p_{\mathcal{X}}(\mathbf{x}) \mathbb{I}[y \neq f(\mathbf{x})] d\mathbf{x} \\
&= \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{y\in\{0,1\}} \int p_{\mathcal{Y}}(y|\mathbf{x}) p_{\mathcal{X}}(\mathbf{x}) \Big( f(\mathbf{x})(1-y) + (1-f(\mathbf{x}))y \Big) d\mathbf{x} \\
&= \operatorname*{argmin}_{f\in\mathcal{F}} \int \Big( p_{\mathcal{Y}}(y=1|\mathbf{x})(1-f(\mathbf{x})) + p_{\mathcal{Y}}(y=0|\mathbf{x}) f(\mathbf{x}) \Big) p_{\mathcal{X}}(\mathbf{x}) d\mathbf{x} \\
&= \operatorname*{argmin}_{f\in\mathcal{F}} \int \Big( p_{\mathcal{Y}}(y=1|\mathbf{x})(1-f(\mathbf{x})) + (1-p_{\mathcal{Y}}(y=1|\mathbf{x})) f(\mathbf{x}) \Big) p_{\mathcal{X}}(\mathbf{x}) d\mathbf{x} \\
&= \operatorname*{argmin}_{f\in\mathcal{F}} \int \Big( p_{\mathcal{Y}}(y=1|\mathbf{x}) + f(\mathbf{x})(1-2p_{\mathcal{Y}}(y=1|\mathbf{x})) \Big) p_{\mathcal{X}}(\mathbf{x}) d\mathbf{x}
\end{aligned}
$$

# Optimisation Problem

- We need to find a function, $f^*$, which is optimal for all $\mathbf{x} \in \text{dom}(\mathcal{X})$:

$$f^*(\mathbf{x}) = \underset{f(\mathbf{x})}{\text{argmin}} \left(p_{\mathcal{Y}}(y=1|\mathbf{x}) + f(\mathbf{x})(1 - 2p_{\mathcal{Y}}(y=1|\mathbf{x}))\right)$$
$$= \underset{f(\mathbf{x})}{\text{argmin}}\, \mathrm{M}(\mathbf{x})$$

- This is a **discrete optimisation problem**:
  - if: $f(\mathbf{x}) = 1$ then: $\mathrm{M}(\mathbf{x}) = 1 - p_{\mathcal{Y}}(y=1|\mathbf{x})$
  - if: $f(\mathbf{x}) = 0$ then: $\mathrm{M}(\mathbf{x}) = p_{\mathcal{Y}}(y=1|\mathbf{x})$

- This means that if $f^*(\mathbf{x}) = 1$ optimality implies:

$$1 - p_{\mathcal{Y}}(y=1|\mathbf{x}) \leqslant p_{\mathcal{Y}}(y=1|\mathbf{x})$$
$$p_{\mathcal{Y}}(y=1|\mathbf{x}) \geqslant 0.5$$

# Bayes Optimal Classifier

- So the generalisation minimiser for the **Misclassification Loss** can be specified entirely in term of the **posterior distribution**:

$$f^*(\mathbf{x}) = \begin{cases} 1 & \text{if} \quad p_{\mathcal{Y}}(y = 1|\mathbf{x}) \geqslant 0.5 \\ 0 & \text{if} \quad p_{\mathcal{Y}}(y = 1|\mathbf{x}) < 0.5 \end{cases}$$

- It is known as the **Bayes Optimal Classifier**

- Of course, different loss functions will lead to different optimal classifiers...

# Alternative Classification Loss Functions

- Misclassification loss can be expressed in terms of a **loss matrix**:

$$\begin{array}{cc} y=0 & y=1 \\ \left[\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array}\right] \end{array} \begin{array}{l} \\ f(\mathbf{x})=0 \\ f(\mathbf{x})=1 \end{array}$$

- Different loss matrices will lead to different classifiers...

- ...In particular they will lead to classification thresholds on the posterior distribution which differ from 0.5

- For example, cancer diagnosis might exhibit the following matrix:

$$\begin{array}{cc} y=0 & y=1 \\ \left[\begin{array}{cc} 0 & 1,000 \\ 1 & 0 \end{array}\right] \end{array} \begin{array}{l} \\ f(\mathbf{x})=0 \\ f(\mathbf{x})=1 \end{array}$$

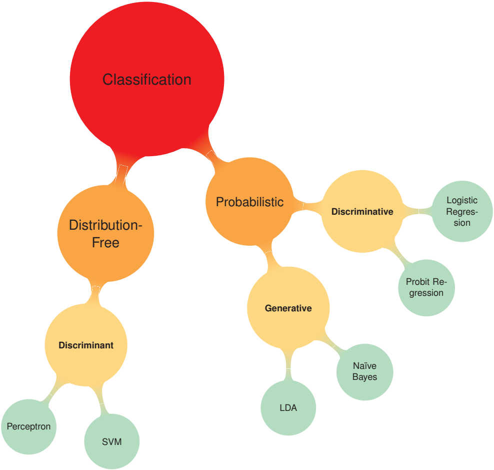
# Probabilistic Classifier

- Characterisation of $f^*$ in terms of $p_{\mathcal{Y}}(y = 1|\mathbf{x})$ suggests a classification paradigm:
- **Probabilistic Classification**
  - The classification problem reduces to an inference problem in which we must learn the posterior output class probability, $p_{\mathcal{Y}}(y = 1|\mathbf{x})$
  - Here $p_{\mathcal{Y}}(y = 1|\mathbf{x})$ characterises an **inhomogeneous Bernoulli distribution**
  - We must learn a Bernoulli distribution for each $\mathbf{x} \in \mathrm{dom}(\mathcal{X})$

# Probabilistic Classifier

- We contrast this with an alternative paradigm:
- **Distribution-Free Classification**
  - Here we seek to learn the classification boundary (equivalently $f^*$) directly, without resorting to probabilistic inference
  - An example is the PAC approach where we seek to approximate $\mathbb{E}_{\mathcal{D}}\left[\mathbb{I}[\mathcal{Y} \neq f(\mathcal{X})]\right]$ without reference to any explicit pdf and then to optimise this new quantity in order to learn $f^*$

# Classification Approaches



# Generative Classification

- We note **Bayes' Theorem**:

$$p_{\mathcal{Y}}(y|\mathbf{x}) = \frac{p_{\mathcal{X}}(\mathbf{x}|y)p_{\mathcal{Y}}(y)}{p_{\mathcal{X}}(\mathbf{x})}$$
$$= \frac{p_{\mathcal{X}}(\mathbf{x}|y)p_{\mathcal{Y}}(y)}{\sum_{y\in\{0,1\}} p_{\mathcal{X}}(\mathbf{x}|y)p_{\mathcal{Y}}(y)}$$

- Learn $p_{\mathcal{Y}}(y|\mathbf{x})$ indirectly by inferring $p_{\mathcal{X}}(\mathbf{x}|y)$ and $p_{\mathcal{Y}}(y)$ for each class separately
- Most demanding approach in terms of number of parameters to learn
- Allows us to learn $p_{\mathcal{X}}(\mathbf{x})$ which can be useful in **novelty detection**

# Discriminative Classification

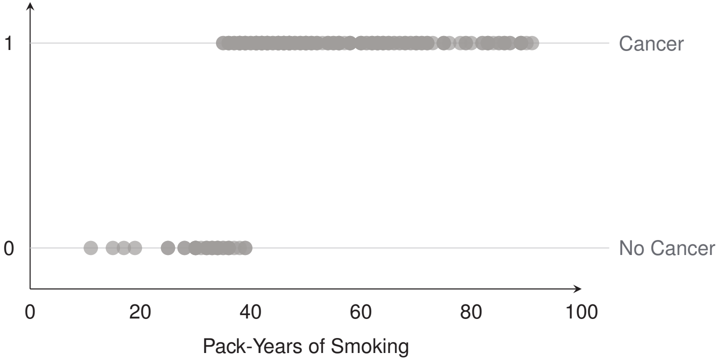
- Attempts to learn $p_{\mathcal{Y}}(y|\mathbf{x})$ directly
- Less demanding in terms of number of parameters to learn

# Discriminant Classification

- As mentioned, this does not seek to infer the pdf at all.
  - Instead it seeks to learn $f^*$ directly
- Least demanding in terms of number of parameters to learn
- Inflexible - requires us to run algorithm afresh for changes of the loss function
  - c.f. probabilistic approaches, which can be used to update $f^*$ trivially
- Does not deal well with **class imbalance**

# Lecture Overview

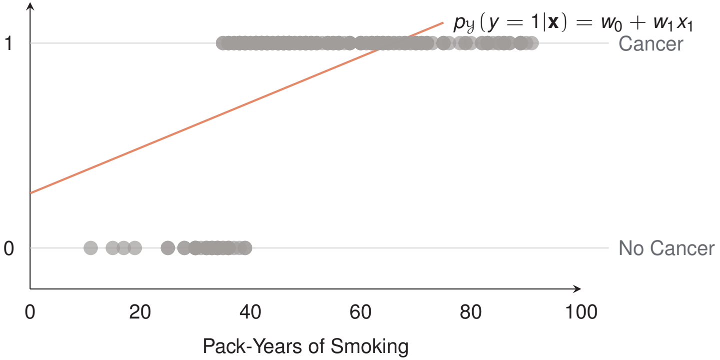
# Throat Cancer Prediction



# Discriminative Classification

- Let us focus on learning the inhomogeneous Bernoulli distribution $p_{\mathcal{y}}(y = 1|\mathbf{x})$ directly
- Can we model $p_{\mathcal{y}}(y = 1|\mathbf{x})$ as a linear function of $\mathbf{x}$, i.e. $\mathbf{w} \cdot \mathbf{x}$
- In other words can we just use Linear Regression?

# Throat Cancer Prediction: Linear Regression

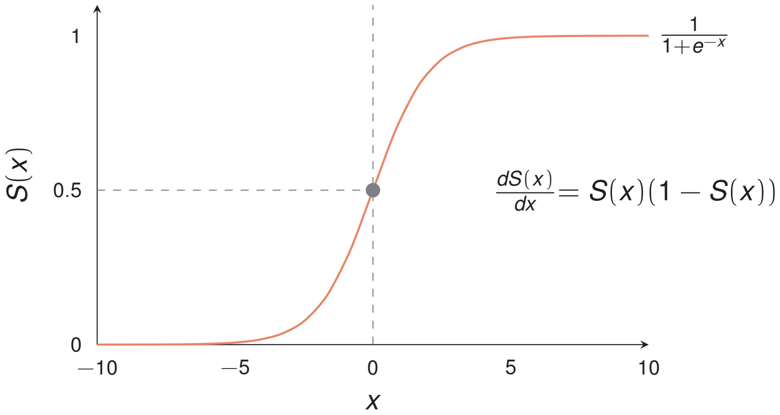


# Discriminative Classification: Linear Regression

- Can we just use Linear Regression?
- No!
  - $p_{\mathcal{Y}}(y|\mathbf{x})$ must lie in the range $[0, 1]$, while linear functions are unbounded
  - We need to learn a function that will **squash** the output of our model into $[0, 1]$
  - One such function is the **logistic sigmoid**, $S$:

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

# Logistic Sigmoid



# Logistic Regression Model

- So we could attempt to model $p_{\mathcal{Y}}(y = 1|\mathbf{x})$ as:

$$p_{\mathcal{Y}}(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}\cdot\mathbf{x}}}$$

- In other words our representation, $\mathcal{F}$, becomes:

$$\mathcal{F} = \left\{ f_{\mathbf{w}}(\mathbf{x}) = \mathbb{I}\left[p_{\mathcal{Y}}(y = 1|\mathbf{x}) \geqslant 0.5\right] \middle| p_{\mathcal{Y}}(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}\cdot\mathbf{x}}}, \mathbf{w} \in \mathbb{R}^{m+1} \right\}$$

- This is the **Logistic Regression Model**

# Odds Ratio

- Re-arranging:

$$
\begin{aligned}
p_{\mathrm{y}}(y=1|\mathbf{x}) &= \frac{1}{1+e^{-\mathbf{w}\cdot\mathbf{x}}} \\
e^{-\mathbf{w}\cdot\mathbf{x}} &= \frac{1}{p_{\mathrm{y}}(y=1|\mathbf{x})} - 1 \\
\mathbf{w}\cdot\mathbf{x} &= \ln\left(\frac{p_{\mathrm{y}}(y=1|\mathbf{x})}{1-p_{\mathrm{y}}(y=1|\mathbf{x})}\right) \\
&= \operatorname{logit}(p_{\mathrm{y}}(y=1|\mathbf{x}))
\end{aligned}
$$

- Here $\frac{p_{\mathrm{y}}(y=1|\mathbf{x})}{1-p_{\mathrm{y}}(y=1|\mathbf{x})}$ is the **odds ratio**

# Linear Discriminant

- If $f_{\mathbf{w}}(\mathbf{x}) = 1$ then $p_{\mathcal{Y}}(y = 1|\mathbf{x}) \geqslant 0.5$, $\frac{p_{\mathcal{Y}}(y=1|\mathbf{x})}{1-p_{\mathcal{Y}}(y=1|\mathbf{x})} \geqslant 1$, $\mathbf{w} \cdot \mathbf{x} \geqslant 0$
- If $f_{\mathbf{w}}(\mathbf{x}) = 0$ then $p_{\mathcal{Y}}(y = 1|\mathbf{x}) < 0.5$, $\frac{p_{\mathcal{Y}}(y=1|\mathbf{x})}{1-p_{\mathcal{Y}}(y=1|\mathbf{x})} < 1$, $\mathbf{w} \cdot \mathbf{x} < 0$
- And $\mathbf{w} \cdot \mathbf{x} = 0$ defines a **linear discriminant** separating hyperplane

# Evaluation

- How should we learn the parameters $\mathbf{w}$?
- This is equivalent to asking how we should learn the distribution $p_{\mathcal{Y}}(y|\mathbf{x})$
- From our earlier work on Probability and Point Estimation recall that we may adopt a number of different approaches
- The most common one is to assume a **frequentist** setting and use **maximum likelihood estimation**

# Evaluation

- Let us construct our log-likelihood function given Bernoulli outcomes:

$$
\begin{aligned}
\ln(\mathrm{L}(\mathbf{w})) &= \ln\left(\prod_{i=1}^{n} p_{\mathrm{y}}(y^{(i)}|\mathbf{x}^{(i)})\right) \\
&= \sum_{i=1}^{n} \ln\left(p_{\mathrm{y}}(y^{(i)}|\mathbf{x}^{(i)})\right) \\
&= \sum_{i=1}^{n} y^{(i)} \ln\left(p_{\mathrm{y}}(y^{(i)}=1|\mathbf{x}^{(i)})\right) \\
&\qquad + (1-y^{(i)}) \ln\left(p_{\mathrm{y}}(y^{(i)}=0|\mathbf{x}^{(i)})\right)
\end{aligned}
$$

- This expression is known as the **Cross-Entropy** loss function

# Evaluation

- Previous two slides are true in general, we now specialise to the logistic regression model by substituting the following:

$$p_{\mathcal{Y}}(y=1|\mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}\cdot\mathbf{x}}} \qquad p_{\mathcal{Y}}(y=0|\mathbf{x}) = \frac{e^{-\mathbf{w}\cdot\mathbf{x}}}{1+e^{-\mathbf{w}\cdot\mathbf{x}}} = \frac{1}{1+e^{\mathbf{w}\cdot\mathbf{x}}}$$

- Into the log-likelihood function:

$$\begin{aligned}\ln(\mathrm{L}(\mathbf{w})) &= \sum_{i=1}^{n} \ln(p_{\mathcal{Y}}(y^{(i)}=0|\mathbf{x}^{(i)})) + y^{(i)} \ln \frac{p_{\mathcal{Y}}(y^{(i)}=1|\mathbf{x}^{(i)})}{p_{\mathcal{Y}}(y^{(i)}=0|\mathbf{x}^{(i)})} \\ &= \sum_{i=1}^{n} y^{(i)}\mathbf{w}\cdot\mathbf{x}^{(i)} - \ln(1+e^{\mathbf{w}\cdot\mathbf{x}^{(i)}})\end{aligned}$$

# Optimisation

- Thus we seek $\mathbf{w}_{\text{MLE}}$, such that:

$$\mathbf{w}_{\text{MLE}} = \underset{\mathbf{w}}{\text{argmin}} \sum_{i=1}^{n} \ln(1 + e^{\mathbf{w}\cdot\mathbf{x}^{(i)}}) - y^{(i)}\mathbf{w} \cdot \mathbf{x}^{(i)}$$

# Motivation

- What leads us to adopt logistic regression as our model?
- After all, plenty of other ‘squashing’ functions exist, apart from the sigmoid
- For example the **cumulative distribution function** of the **standard normal distribution**

# Motivation 1: Additive Logistic Noise Latent Variable Model

- Let us assume that our data is generated by some hidden (**latent**) random variable, $\mathcal{Y}^*$, such that:

$$y^* = \mathbf{w} \cdot \mathbf{x} + \varepsilon$$

Where $\varepsilon$ is an outcome associated with a random variable $\epsilon$, distributed according to a Logistic Distribution, $\varepsilon \sim \text{Logistic}(0, 1)$; and where $y^*$ are the outcomes associated with $\mathcal{Y}^*$

# Logistic Distribution

- Let $\epsilon$ be a continuous random variable, taking values $\varepsilon \in \mathbb{R}$, with a **Logistic** distribution:

$$\varepsilon \sim \text{Logistic}(\mu, s) \qquad \text{where:} \qquad \mu \in \mathbb{R}, s > 0$$

- This has a characteristic pdf, $f_\epsilon$:

$$f_\epsilon(\varepsilon; \mu, s) = \frac{e^{-\frac{\varepsilon - \mu}{s}}}{s\left(1 + e^{-\frac{\varepsilon - \mu}{s}}\right)^2}$$

$$\mathbb{E}_{\mathcal{D}}[\epsilon] = \mu$$

$$\mathbb{P}(\epsilon < z) = \frac{1}{1 + e^{-\frac{z - \mu}{s}}}$$

# Motivation 1: Additive Logistic Noise Latent Variable Model

- We link a latent variable outcome, $y^*$, to an output, $y$, by assuming the following classification model:

$$y^{(i)} = \begin{cases} 1 & \text{if} \quad y^{*(i)} \geqslant 0 \\ 0 & \text{if} \quad y^{*(i)} < 0 \end{cases}$$

# Motivation 1: Additive Logistic Noise Latent Variable Model

- Then:

$$p_{\mathcal{Y}}(y^{(i)} = 1|\mathbf{x}^{(i)}) = \mathbb{P}(\mathcal{Y}^{*(i)} \geqslant 0|\mathbf{x}^{(i)})$$

# Motivation 1: Additive Logistic Noise Latent Variable Model

- Then:

$$
\begin{aligned}
p_{\mathcal{Y}}(y^{(i)} = 1|\mathbf{x}^{(i)}) &= \mathbb{P}(\mathcal{Y}^{*(i)} \geqslant 0|\mathbf{x}^{(i)}) \\
&= \mathbb{P}(\mathbf{w} \cdot \mathbf{x}^{(i)} + \epsilon^{(i)} \geqslant 0)
\end{aligned}
$$

# Motivation 1: Additive Logistic Noise Latent Variable Model

- Then:

$$
\begin{aligned}
p_{\mathcal{Y}}(y^{(i)} = 1|\mathbf{x}^{(i)}) &= \mathbb{P}(\mathcal{Y}^{*(i)} \geqslant 0|\mathbf{x}^{(i)}) \\
&= \mathbb{P}(\mathbf{w} \cdot \mathbf{x}^{(i)} + \epsilon^{(i)} \geqslant 0) \\
&= \mathbb{P}(\epsilon^{(i)} \geqslant -\mathbf{w} \cdot \mathbf{x}^{(i)})
\end{aligned}
$$

# Motivation 1: Additive Logistic Noise Latent Variable Model

- Then:

$$\begin{aligned} p_{\mathcal{Y}}(y^{(i)} = 1|\mathbf{x}^{(i)}) &= \mathbb{P}(\mathcal{Y}^{*(i)} \geqslant 0|\mathbf{x}^{(i)}) \\ &= \mathbb{P}(\mathbf{w} \cdot \mathbf{x}^{(i)} + \epsilon^{(i)} \geqslant 0) \\ &= \mathbb{P}(\epsilon^{(i)} \geqslant -\mathbf{w} \cdot \mathbf{x}^{(i)}) \\ &= 1 - \mathbb{P}(\epsilon^{(i)} < -\mathbf{w} \cdot \mathbf{x}^{(i)}) \end{aligned}$$

# Motivation 1: Additive Logistic Noise Latent Variable Model

- Then:

$$
\begin{aligned}
p_{\mathcal{Y}}(y^{(i)} = 1|\mathbf{x}^{(i)}) &= \mathbb{P}(\mathcal{Y}^{*(i)} \geqslant 0|\mathbf{x}^{(i)}) \\
&= \mathbb{P}(\mathbf{w} \cdot \mathbf{x}^{(i)} + \epsilon^{(i)} \geqslant 0) \\
&= \mathbb{P}(\epsilon^{(i)} \geqslant -\mathbf{w} \cdot \mathbf{x}^{(i)}) \\
&= 1 - \mathbb{P}(\epsilon^{(i)} < -\mathbf{w} \cdot \mathbf{x}^{(i)}) \\
&= 1 - \frac{1}{1 + e^{\mathbf{w} \cdot \mathbf{x}^{(i)}}}
\end{aligned}
$$

# Motivation 1: Additive Logistic Noise Latent Variable Model

- Then:

$$
\begin{aligned}
p_{\mathcal{Y}}(y^{(i)} = 1|\mathbf{x}^{(i)}) &= \mathbb{P}(\mathcal{Y}^{*(i)} \geqslant 0|\mathbf{x}^{(i)}) \\
&= \mathbb{P}(\mathbf{w} \cdot \mathbf{x}^{(i)} + \epsilon^{(i)} \geqslant 0) \\
&= \mathbb{P}(\epsilon^{(i)} \geqslant -\mathbf{w} \cdot \mathbf{x}^{(i)}) \\
&= 1 - \mathbb{P}(\epsilon^{(i)} < -\mathbf{w} \cdot \mathbf{x}^{(i)}) \\
&= 1 - \frac{1}{1 + e^{\mathbf{w} \cdot \mathbf{x}^{(i)}}} \\
&= \frac{e^{\mathbf{w} \cdot \mathbf{x}^{(i)}}}{1 + e^{\mathbf{w} \cdot \mathbf{x}^{(i)}}}
\end{aligned}
$$

# Motivation 1: Additive Logistic Noise Latent Variable Model

- Then:

$$
\begin{aligned}
p_{\mathcal{Y}}(y^{(i)} = 1|\mathbf{x}^{(i)}) &= \mathbb{P}(\mathcal{Y}^{*(i)} \geqslant 0|\mathbf{x}^{(i)}) \\
&= \mathbb{P}(\mathbf{w} \cdot \mathbf{x}^{(i)} + \epsilon^{(i)} \geqslant 0) \\
&= \mathbb{P}(\epsilon^{(i)} \geqslant -\mathbf{w} \cdot \mathbf{x}^{(i)}) \\
&= 1 - \mathbb{P}(\epsilon^{(i)} < -\mathbf{w} \cdot \mathbf{x}^{(i)}) \\
&= 1 - \frac{1}{1 + e^{\mathbf{w} \cdot \mathbf{x}^{(i)}}} \\
&= \frac{e^{\mathbf{w} \cdot \mathbf{x}^{(i)}}}{1 + e^{\mathbf{w} \cdot \mathbf{x}^{(i)}}} \\
&= \frac{1}{1 + e^{-\mathbf{w} \cdot \mathbf{x}^{(i)}}}
\end{aligned}
$$

# Motivation 1: Additive Logistic Noise Latent Variable Model

- Then:

$$
\begin{aligned}
p_{\mathcal{Y}}(y^{(i)} = 1|\mathbf{x}^{(i)}) &= \mathbb{P}(\mathcal{Y}^{*(i)} \geqslant 0|\mathbf{x}^{(i)}) \\
&= \mathbb{P}(\mathbf{w} \cdot \mathbf{x}^{(i)} + \epsilon^{(i)} \geqslant 0) \\
&= \mathbb{P}(\epsilon^{(i)} \geqslant -\mathbf{w} \cdot \mathbf{x}^{(i)}) \\
&= 1 - \mathbb{P}(\epsilon^{(i)} < -\mathbf{w} \cdot \mathbf{x}^{(i)}) \\
&= 1 - \frac{1}{1 + e^{\mathbf{w} \cdot \mathbf{x}^{(i)}}} \\
&= \frac{e^{\mathbf{w} \cdot \mathbf{x}^{(i)}}}{1 + e^{\mathbf{w} \cdot \mathbf{x}^{(i)}}} \\
&= \frac{1}{1 + e^{-\mathbf{w} \cdot \mathbf{x}^{(i)}}}
\end{aligned}
$$

- ...Which is the logistic regression model

# Motivation 2

- Logistic regression leads to an optimisation problem which itself has directly attractive properties, it is:
  - **Smooth**
  - **Convex**

- We will see that we have turned a **non-convex problem** - trying to optimise $\mathbb{E}_{\mathcal{D}}\left[\mathbb{I}[f(\mathcal{X}) \neq \mathcal{Y}]\right]$ - into a **convex problem**

- We do this by shifting our focus to probabilistic inference, and selecting a tractable form for $p_{\mathcal{X},\mathcal{Y}}$

- In practice it works well as a classifier (although it does not always get the probability prediction right)

# Recap

- **Representation:**

$$\mathcal{F} = \left\{ f_{\mathbf{w}}(\mathbf{x}) = \mathbb{I}\left[p_{\mathcal{Y}}(y = 1|\mathbf{x}) \geqslant 0.5\right] \middle| p_{\mathcal{Y}}(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}\cdot\mathbf{x}}}, \mathbf{w} \in \mathbb{R}^{m+1} \right\}$$

- **Evaluation:**

$$\sum_{i=1}^{n} y^{(i)} \mathbf{w} \cdot \mathbf{x}^{(i)} - \ln\left(1 + e^{\mathbf{w}\cdot\mathbf{x}^{(i)}}\right) \qquad \textbf{i.e. the log-likelihood}$$

- **Optimisation:**

$$\mathbf{w}_{\mathrm{MLE}} = \operatorname*{argmin}_{\mathbf{w}} \sum_{i=1}^{n} \ln\left(1 + e^{\mathbf{w}\cdot\mathbf{x}^{(i)}}\right) - y^{(i)} \mathbf{w} \cdot \mathbf{x}^{(i)}$$

equivalent to a maximisation of the log-likelihood

# Optimisation

- We seek:

$$\mathbf{w}_{\text{MLE}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left( \sum_{i=1}^{n} \ln(1 + e^{\mathbf{w}\cdot\mathbf{x}^{(i)}}) - y^{(i)}\mathbf{w}\cdot\mathbf{x}^{(i)} \right)$$

- But this has no analytic solution...
- ...However, we can apply a numerical technique to optimise, such as **gradient descent**
- But we should really check for **convexity**

# Convexity

- The sum of convex functions is also a convex function (think about this)
- So if we can prove that $-y\mathbf{w}\cdot\mathbf{x}$ and $\ln(1+e^{\mathbf{w}\cdot\mathbf{x}})$ are convex, then we have proved convexity of our objective
- Consider $-y\mathbf{w}\cdot\mathbf{x}$:
$$\nabla_{\mathbf{w}}(-y\mathbf{w}\cdot\mathbf{x}) = -y\mathbf{x}$$
- Taking second derivatives:
$$\begin{aligned} {\nabla^2}_{\mathbf{w}}(-y\mathbf{w}\cdot\mathbf{x}) &= \mathbf{0} \\ \implies \quad \mathcal{H} &= \mathbf{0} \qquad \text{which demonstrates convexity} \end{aligned}$$

# Convexity

- Consider $\ln(1 + e^{\mathbf{w}\cdot\mathbf{x}})$:

$$\nabla_{\mathbf{w}}(\ln(1 + e^{\mathbf{w}\cdot\mathbf{x}})) = \frac{e^{\mathbf{w}\cdot\mathbf{x}}}{1 + e^{\mathbf{w}\cdot\mathbf{x}}}\mathbf{x}$$

- So:

$$\frac{\partial(\ln(1 + e^{\mathbf{w}\cdot\mathbf{x}}))}{\partial w_i} = \frac{e^{\mathbf{w}\cdot\mathbf{x}}}{1 + e^{\mathbf{w}\cdot\mathbf{x}}}x_i$$

# Convexity

- Thus:

$$\frac{\partial^2(\ln(1+e^{\mathbf{w}\cdot\mathbf{x}}))}{\partial w_i \partial w_j} = \frac{\partial(e^{\mathbf{w}\cdot\mathbf{x}})}{\partial w_j}\frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} + \frac{\partial\left(\frac{1}{1+e^{\mathbf{w}\cdot\mathbf{x}}}\right)}{\partial w_j} e^{\mathbf{w}\cdot\mathbf{x}} x_i$$

# Convexity

- Thus:

$$\frac{\partial^2(\ln(1+e^{\mathbf{w}\cdot\mathbf{x}}))}{\partial w_i \partial w_j} = \frac{\partial(e^{\mathbf{w}\cdot\mathbf{x}})}{\partial w_j}\frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} + \frac{\partial\left(\frac{1}{1+e^{\mathbf{w}\cdot\mathbf{x}}}\right)}{\partial w_j} e^{\mathbf{w}\cdot\mathbf{x}} x_i$$

$$= e^{\mathbf{w}\cdot\mathbf{x}}\frac{\partial(\mathbf{w}\cdot\mathbf{x})}{\partial w_j}\frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} - \frac{1}{(1+e^{\mathbf{w}\cdot\mathbf{x}})^2}\frac{\partial(1+e^{\mathbf{w}\cdot\mathbf{x}})}{\partial w_j} e^{\mathbf{w}\cdot\mathbf{x}} x_i$$

# Convexity

- Thus:

$$
\begin{aligned}
\frac{\partial^2(\ln(1+e^{\mathbf{w}\cdot\mathbf{x}}))}{\partial w_i \partial w_j} &= \frac{\partial(e^{\mathbf{w}\cdot\mathbf{x}})}{\partial w_j}\frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} + \frac{\partial\left(\frac{1}{1+e^{\mathbf{w}\cdot\mathbf{x}}}\right)}{\partial w_j} e^{\mathbf{w}\cdot\mathbf{x}} x_i \\
&= e^{\mathbf{w}\cdot\mathbf{x}}\frac{\partial(\mathbf{w}\cdot\mathbf{x})}{\partial w_j}\frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} - \frac{1}{(1+e^{\mathbf{w}\cdot\mathbf{x}})^2}\frac{\partial(1+e^{\mathbf{w}\cdot\mathbf{x}})}{\partial w_j} e^{\mathbf{w}\cdot\mathbf{x}} x_i \\
&= e^{\mathbf{w}\cdot\mathbf{x}} x_j \frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} - \frac{1}{(1+e^{\mathbf{w}\cdot\mathbf{x}})^2} e^{\mathbf{w}\cdot\mathbf{x}}\frac{\partial(\mathbf{w}\cdot\mathbf{x})}{\partial w_j} e^{\mathbf{w}\cdot\mathbf{x}} x_i
\end{aligned}
$$

# Convexity

- Thus:

$$\begin{aligned}
\frac{\partial^2(\ln(1+e^{\mathbf{w}\cdot\mathbf{x}}))}{\partial w_i \partial w_j} &= \frac{\partial(e^{\mathbf{w}\cdot\mathbf{x}})}{\partial w_j}\frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} + \frac{\partial\left(\frac{1}{1+e^{\mathbf{w}\cdot\mathbf{x}}}\right)}{\partial w_j} e^{\mathbf{w}\cdot\mathbf{x}} x_i \\
&= e^{\mathbf{w}\cdot\mathbf{x}}\frac{\partial(\mathbf{w}\cdot\mathbf{x})}{\partial w_j}\frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} - \frac{1}{(1+e^{\mathbf{w}\cdot\mathbf{x}})^2}\frac{\partial(1+e^{\mathbf{w}\cdot\mathbf{x}})}{\partial w_j} e^{\mathbf{w}\cdot\mathbf{x}} x_i \\
&= e^{\mathbf{w}\cdot\mathbf{x}} x_j \frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} - \frac{1}{(1+e^{\mathbf{w}\cdot\mathbf{x}})^2} e^{\mathbf{w}\cdot\mathbf{x}} \frac{\partial(\mathbf{w}\cdot\mathbf{x})}{\partial w_j} e^{\mathbf{w}\cdot\mathbf{x}} x_i \\
&= e^{\mathbf{w}\cdot\mathbf{x}} x_j \frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} - \frac{1}{(1+e^{\mathbf{w}\cdot\mathbf{x}})^2} e^{\mathbf{w}\cdot\mathbf{x}} x_j e^{\mathbf{w}\cdot\mathbf{x}} x_i
\end{aligned}$$

# Convexity

- Thus:

$$
\begin{aligned}
\frac{\partial^2(\ln(1+e^{\mathbf{w}\cdot\mathbf{x}}))}{\partial w_i \partial w_j} &= \frac{\partial(e^{\mathbf{w}\cdot\mathbf{x}})}{\partial w_j}\frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} + \frac{\partial\left(\frac{1}{1+e^{\mathbf{w}\cdot\mathbf{x}}}\right)}{\partial w_j} e^{\mathbf{w}\cdot\mathbf{x}} x_i \\
&= e^{\mathbf{w}\cdot\mathbf{x}}\frac{\partial(\mathbf{w}\cdot\mathbf{x})}{\partial w_j}\frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} - \frac{1}{(1+e^{\mathbf{w}\cdot\mathbf{x}})^2}\frac{\partial(1+e^{\mathbf{w}\cdot\mathbf{x}})}{\partial w_j} e^{\mathbf{w}\cdot\mathbf{x}} x_i \\
&= e^{\mathbf{w}\cdot\mathbf{x}} x_j \frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} - \frac{1}{(1+e^{\mathbf{w}\cdot\mathbf{x}})^2} e^{\mathbf{w}\cdot\mathbf{x}} \frac{\partial(\mathbf{w}\cdot\mathbf{x})}{\partial w_j} e^{\mathbf{w}\cdot\mathbf{x}} x_i \\
&= e^{\mathbf{w}\cdot\mathbf{x}} x_j \frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} - \frac{1}{(1+e^{\mathbf{w}\cdot\mathbf{x}})^2} e^{\mathbf{w}\cdot\mathbf{x}} x_j e^{\mathbf{w}\cdot\mathbf{x}} x_i \\
&= \frac{e^{\mathbf{w}\cdot\mathbf{x}}}{1+e^{\mathbf{w}\cdot\mathbf{x}}}\left(1 - \frac{e^{\mathbf{w}\cdot\mathbf{x}}}{1+e^{\mathbf{w}\cdot\mathbf{x}}}\right) x_i x_j
\end{aligned}
$$

# Convexity

- Thus:

$$\begin{aligned}
\frac{\partial^2(\ln(1+e^{\mathbf{w}\cdot\mathbf{x}}))}{\partial w_i \partial w_j} &= \frac{\partial(e^{\mathbf{w}\cdot\mathbf{x}})}{\partial w_j}\frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} + \frac{\partial\left(\frac{1}{1+e^{\mathbf{w}\cdot\mathbf{x}}}\right)}{\partial w_j} e^{\mathbf{w}\cdot\mathbf{x}} x_i \\
&= e^{\mathbf{w}\cdot\mathbf{x}}\frac{\partial(\mathbf{w}\cdot\mathbf{x})}{\partial w_j}\frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} - \frac{1}{(1+e^{\mathbf{w}\cdot\mathbf{x}})^2}\frac{\partial(1+e^{\mathbf{w}\cdot\mathbf{x}})}{\partial w_j} e^{\mathbf{w}\cdot\mathbf{x}} x_i \\
&= e^{\mathbf{w}\cdot\mathbf{x}} x_j \frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} - \frac{1}{(1+e^{\mathbf{w}\cdot\mathbf{x}})^2} e^{\mathbf{w}\cdot\mathbf{x}}\frac{\partial(\mathbf{w}\cdot\mathbf{x})}{\partial w_j} e^{\mathbf{w}\cdot\mathbf{x}} x_i \\
&= e^{\mathbf{w}\cdot\mathbf{x}} x_j \frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} - \frac{1}{(1+e^{\mathbf{w}\cdot\mathbf{x}})^2} e^{\mathbf{w}\cdot\mathbf{x}} x_j e^{\mathbf{w}\cdot\mathbf{x}} x_i \\
&= \frac{e^{\mathbf{w}\cdot\mathbf{x}}}{1+e^{\mathbf{w}\cdot\mathbf{x}}}\left(1 - \frac{e^{\mathbf{w}\cdot\mathbf{x}}}{1+e^{\mathbf{w}\cdot\mathbf{x}}}\right) x_i x_j \\
&= \frac{e^{\mathbf{w}\cdot\mathbf{x}}}{1+e^{\mathbf{w}\cdot\mathbf{x}}}\left(\frac{1}{1+e^{\mathbf{w}\cdot\mathbf{x}}}\right) x_i x_j
\end{aligned}$$

# Convexity

- Thus:

$$
\begin{aligned}
\frac{\partial^2(\ln(1+e^{\mathbf{w}\cdot\mathbf{x}}))}{\partial w_i \partial w_j} &= \frac{\partial(e^{\mathbf{w}\cdot\mathbf{x}})}{\partial w_j}\frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} + \frac{\partial\left(\frac{1}{1+e^{\mathbf{w}\cdot\mathbf{x}}}\right)}{\partial w_j} e^{\mathbf{w}\cdot\mathbf{x}} x_i \\
&= e^{\mathbf{w}\cdot\mathbf{x}}\frac{\partial(\mathbf{w}\cdot\mathbf{x})}{\partial w_j}\frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} - \frac{1}{(1+e^{\mathbf{w}\cdot\mathbf{x}})^2}\frac{\partial(1+e^{\mathbf{w}\cdot\mathbf{x}})}{\partial w_j} e^{\mathbf{w}\cdot\mathbf{x}} x_i \\
&= e^{\mathbf{w}\cdot\mathbf{x}} x_j \frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} - \frac{1}{(1+e^{\mathbf{w}\cdot\mathbf{x}})^2} e^{\mathbf{w}\cdot\mathbf{x}} \frac{\partial(\mathbf{w}\cdot\mathbf{x})}{\partial w_j} e^{\mathbf{w}\cdot\mathbf{x}} x_i \\
&= e^{\mathbf{w}\cdot\mathbf{x}} x_j \frac{x_i}{1+e^{\mathbf{w}\cdot\mathbf{x}}} - \frac{1}{(1+e^{\mathbf{w}\cdot\mathbf{x}})^2} e^{\mathbf{w}\cdot\mathbf{x}} x_j e^{\mathbf{w}\cdot\mathbf{x}} x_i \\
&= \frac{e^{\mathbf{w}\cdot\mathbf{x}}}{1+e^{\mathbf{w}\cdot\mathbf{x}}}\left(1 - \frac{e^{\mathbf{w}\cdot\mathbf{x}}}{1+e^{\mathbf{w}\cdot\mathbf{x}}}\right) x_i x_j \\
&= \frac{e^{\mathbf{w}\cdot\mathbf{x}}}{1+e^{\mathbf{w}\cdot\mathbf{x}}}\left(\frac{1}{1+e^{\mathbf{w}\cdot\mathbf{x}}}\right) x_i x_j \\
&= \frac{e^{\mathbf{w}\cdot\mathbf{x}}}{(1+e^{\mathbf{w}\cdot\mathbf{x}})^2} x_i x_j
\end{aligned}
$$

## Convexity

- Now, note that:

$$\mathbf{x}\mathbf{x}^T = \begin{bmatrix} x_0x_0 & x_0x_1 & \cdot & \cdot & x_0x_m \\ x_1x_0 & x_1x_1 & \cdot & \cdot & x_1x_m \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x_mx_0 & x_mx_1 & \cdot & \cdot & x_mx_m \end{bmatrix}$$

$$\implies \quad [\mathbf{x}\mathbf{x}^T]_{ij} = x_ix_j$$

- This allows us to express the Hessian for $\ln(1 + e^{\mathbf{w}\cdot\mathbf{x}})$ as:

$$\mathcal{H} = \frac{e^{\mathbf{w}\cdot\mathbf{x}}}{(1 + e^{\mathbf{w}\cdot\mathbf{x}})^2}\mathbf{x}\mathbf{x}^T$$

# Convexity

- Thus:

$$\begin{aligned}
\mathbf{a}^T \mathcal{H} \mathbf{a} &= \frac{e^{\mathbf{w}\cdot\mathbf{x}}}{(1+e^{\mathbf{w}\cdot\mathbf{x}})^2} \mathbf{a}^T \mathbf{x}\mathbf{x}^T \mathbf{a} \\
&= \frac{e^{\mathbf{w}\cdot\mathbf{x}}}{(1+e^{\mathbf{w}\cdot\mathbf{x}})^2} \|\mathbf{a}^T \mathbf{x}\|_2^2 \geqslant 0 \\
\implies \quad & \mathcal{H} \succeq 0
\end{aligned}$$

- So our objective is convex...
- ...And gradient descent will converge to a **global** (but not necessarily **unique**) optimal solution...if one exists

# Perfectly Separable Data

- Consider what happens if our training data is **linearly separable**
- And consider the objective function which we are seeking to maximise, for some $\mathbf{w} = c\widetilde{\mathbf{w}}$, where $c > 0$, which separates the training data:

$$\sum_{i=1}^{n} \ln(1 + e^{c\widetilde{\mathbf{w}}\cdot\mathbf{x}^{(i)}}) - y^{(i)} c\widetilde{\mathbf{w}} \cdot \mathbf{x}^{(i)}$$

# Perfectly Separable Data: Overfitting

- Now let's take the derivative of this with respect to $c$:

$$\sum_{i=1}^{n} \frac{e^{c\widetilde{\mathbf{w}}\cdot\mathbf{x}^{(i)}}}{1+e^{c\widetilde{\mathbf{w}}\cdot\mathbf{x}^{(i)}}}\widetilde{\mathbf{w}}\cdot\mathbf{x}^{(i)} - y^{(i)}\widetilde{\mathbf{w}}\cdot\mathbf{x}^{(i)} = \sum_{i=1}^{n} \widetilde{\mathbf{w}}\cdot\mathbf{x}^{(i)} \left( \frac{e^{c\widetilde{\mathbf{w}}\cdot\mathbf{x}^{(i)}}}{1+e^{c\widetilde{\mathbf{w}}\cdot\mathbf{x}^{(i)}}} - y^{(i)} \right)$$

- But remember that for this **w** the data is well classified, so:
  - If $y^{(i)} = 1$ then $\widetilde{\mathbf{w}} \cdot \mathbf{x}^{(i)} > 0$ and $\frac{e^{c\widetilde{\mathbf{w}}\cdot\mathbf{x}^{(i)}}}{1+e^{c\widetilde{\mathbf{w}}\cdot\mathbf{x}^{(i)}}} - y^{(i)} < 0$
  - If $y^{(i)} = 0$ then $\widetilde{\mathbf{w}} \cdot \mathbf{x}^{(i)} < 0$ and $\frac{e^{c\widetilde{\mathbf{w}}\cdot\mathbf{x}^{(i)}}}{1+e^{c\widetilde{\mathbf{w}}\cdot\mathbf{x}^{(i)}}} - y^{(i)} > 0$
  - In other words $\widetilde{\mathbf{w}} \cdot \mathbf{x}^{(i)} \left( \frac{e^{c\widetilde{\mathbf{w}}\cdot\mathbf{x}^{(i)}}}{1+e^{c\widetilde{\mathbf{w}}\cdot\mathbf{x}^{(i)}}} - y^{(i)} \right) < 0$ for all $i$
- So if the data are linearly separable then there exists a $\widetilde{\mathbf{w}}$ such that the objective has a negative derivative with respect to $c$

# Perfectly Separable Data: Overfitting

- And in this situation gradient descent will cause $c$ to grow without bound
- This drives the sigmoid function to the **Heaviside function** which is infinitely steep at its inflection point
- So this is an example of the non-existence of a (finite) solution to the Logistic Regression optimisation problem
- We can regard it as a case of **overfitting** - our model becomes overly confident about the data and uses very large weights to achieve a ‘perfect fit’.

UCL

# Perfectly Separable Data: Regularisation

- We can circumvent this problem via **regularisation**. For example an $\ell_2$ norm regulariser will lead to the following objective in this situation:

$$\sum_{i=1}^{n} \ln(1 + e^{c\widetilde{\mathbf{w}}\cdot\mathbf{x}^{(i)}}) - y^{(i)} c\widetilde{\mathbf{w}} \cdot \mathbf{x}^{(i)} + \lambda c^2 \|\widetilde{\mathbf{w}}\|_2^2$$

- And the derivative of this with respect to $c$ includes a strictly positive component, $2\lambda c\|\widetilde{\mathbf{w}}\|_2^2$, which allows us to discern a (unique) stationary point with respect to $c$.

# Regularisation

- Of course, this doesn't necessarily mean that other situations will not lead to problems in optimsation...
- ...But actually, it turns out that, apart from the non-regularised, separable case, the optimisation of the Logistic Regression objective does in fact lead to unique solutions [Silvapulle ('81)]
- So the addition of a ridge regulariser serves to ensure that gradient descent will lead to a globally unique solution for any input data

# Multinomial Logistic Regression

- We may consider **multinomial logistic regression** in an equivalent formulation
  - Assume we have *k* classes such that: $y \in \{1, ..., k\}$
- The Bayes Optimal Classifier becomes:

$$f^*(\mathbf{x}) = \underset{y \in \{1,...,k\}}{\operatorname{argmax}} \; p_{\mathcal{Y}}(y|\mathbf{x})$$

# Multinomial Logistic Regression

- The model is now defined using the **softmax function**, and we seek to learn an **inhomogeneous multinomial distribution**:

$$p_{\mathcal{Y}}(y = j|\mathbf{x}) = \frac{e^{\mathbf{w}_j \cdot \mathbf{x}}}{\sum_{j=1}^{k} e^{\mathbf{w}_j \cdot \mathbf{x}}}$$

  Which is characterised by $\{\mathbf{w}_j\}_{j=1}^{k}$

- The maximum likelihood solution will now make use of the **multinomial cross entropy**, and will again result in a convex optimisation problem

# Lecture Overview

# Summary

1. The **Classification** task can be tackled using probabilistic and distribution-free approaches
2. Probabilistic classification can be split into **Discriminative** and **Generative** approaches
3. **Logistic regression** is a well-motivated approach to discriminative classification which leads to a smooth, convex, optimisation problem

In the next lecture we will move on to consider Generative Classification in more detail, and in particular the **Naïve Bayes** algorithm