CHAPTER 10: Logistic Regression
Logistic Regression - Motivation

- Let's now focus on the **binary classification problem** in which
  - y can take on only two values, 0 and 1.
  - x is a vector of real-valued features, \(< x_1 \ldots x_n >\)

- We could approach the classification problem ignoring the fact that y is discrete-valued, and use our old linear regression algorithm to try to predict y given x.
  - However, it doesn't make sense for f(x) to possibly take values larger than 1 or smaller than 0 when we know that y \(\in\) \(\{0, 1\}\).
Since the output must be 0 or 1, we cannot directly use a linear model to estimate $f(x)$.

Furthermore, we would like to $f(x)$ to represent the probability $P(C_1|x)$. Let’s call it $p$.

We will model the log of the odds of the probability $p$ as a linear function of the input $x$.

$$odds = \frac{p}{1 - p}$$

$$\ln \text{ (odds of } p) = \ln \left(\frac{p}{1-p}\right) = w.x$$

This is the logit function. I.e. $\logit(p) = \ln \left(\frac{p}{1-p}\right)$

If there is a 75% chance that it will rain tomorrow, then the odds of it raining tomorrow are 3 to 1. $(\frac{3}{4})/\frac{1}{4}=3/1$. 
We want: \( f(x) = P(C_1 \mid x) = p \)
We will model as: \( \ln \left( \frac{p}{1-p} \right) = w \cdot x \)

- By applying the inverse of the logit function, that is the logistic function, on both sides, we get:
  \[
  \text{logit}^{-1} \left( \ln \left( \frac{p}{1-p} \right) \right) = \text{sigmoid} \left( \ln \left( \frac{p}{1-p} \right) \right) = p
  \]

- Applying it on the RHS as well, we get
  \[
  p = \text{logit}^{-1} \left( w \cdot x \right) = \frac{1}{1 + e^{-w \cdot x}}
  \]

Thus: \( f(x) = \frac{1}{1 + e^{-w \cdot x}} \) and we will interpret it as \( p = P(C_1 \mid x) = P(y=1 \mid x) \)
Odds & Odds Ratios

The odds has a range of 0 to $\infty$ with values:

- greater than 1 associated with an event being more likely to occur than not to occur and
- values less than 1 associated with an event that is less likely to occur than not occur.

\[
\ln(\text{odds}) = \ln\left(\frac{p}{1-p}\right) = \ln(p) - \ln(1-p)
\]

- The logit is defined as the log of the odds (-$\infty$ to +$\infty$)

  As $\beta.x$ gets really big, $p$ approaches 1
  As $\beta.x$ gets really small, $p$ approaches 0
The Logistic Regression Model

\[ \ln[p/(1-p)] = \beta_0 + \beta_1 X \]

- \( p \) is the probability that the event \( Y \) occurs, \( p(Y=1) \)
  - [range=0 to 1]

- \( p/(1-p) \) is the "odds ratio"
  - [range=0 to \( \infty \)]

- \( \ln[p/(1-p)] \): log odds ratio, or "logit"
  - [range=-\( \infty \) to +\( \infty \)]
We have:

\[ f(x) = \frac{1}{1 + e^{-w \cdot x}} \] and we will interpret it as \( f(x) = P(y=1 \mid x) \)
(in short p)

Thus we have:

\[
\begin{align*}
P(y=1 \mid x) &= f(x) \\
P(y=0 \mid x) &= 1 - f(x)
\end{align*}
\]

Which can be written more compactly by unifying the two rules:

\[
P(y \mid x) = (f(x))^y (1 - f(x))^{1-y} \quad \text{where } y \in \{0, 1\}
\]
1. Calculate $\mathbf{w}^T \mathbf{x}$ and choose $C_1$ if $\mathbf{w}^T \mathbf{x} > 0$, or
2. Calculate $f(\mathbf{x}) = \text{sigmoid}(\mathbf{w}^T \mathbf{x})$ and choose $C_1$ if $f(\mathbf{x}) > 0.5$
Properties

- Linear Decision boundary

- Need for scaling input features:
  - Strictly speaking not needed, but useful in regularized version where we add the weight vector norm (which in turn depends on the scale of the input dimensions) as penalty.
\( P(y \mid x; w) = (f(x))^y (1 - f(x))^{1-y} \)

- Find \( w \) that maximizes the log likelihood of data
  Equivalently, minimizes the negative log likelihood of data

\[
X = \{x^t, y^t\}, \quad y^t \mid x^t \sim \text{Bernoulli}(p)
\]

\[
f(x) = P(y = 1 \mid x) = \frac{1}{1 + \exp[-(w^T x + w_0)]}
\]

\[
l(w, w_0 \mid X) = \prod_t \left( f(x^t)^{y^t} (1 - f(x^t))^{1-y^t} \right)
\]

\[
E = -\log l
\]

\[
E(w, w_0 \mid X) = -\sum_t y^t \log f(x^t) + (1 - y^t) \log (1 - f(x^t))
\]

cross-entropy loss
Cross-entropy loss
Softmax Regression

Multinomial Logistic Regression
MaxEnt Classifier
Softmax Regression

- Softmax regression model generalizes logistic regression to classification problems where the class label $y$ can take on more than two possible values.

- The response variable $y$ can take on any one of $k$ values, so $y \in \{1, 2, \ldots, k\}$. 
Softmax Regression

- Softmax regression model generalizes logistic regression to classification problems where the class label $y$ can take on more than two possible values.

- The response variable $y$ can take on any one of $k$ values, so $y \in \{1, 2, \ldots, K\}$.

$$f(x) = \begin{bmatrix} P(y = 1|x) \\ P(y = 2|x) \\ \vdots \\ P(y = K|x) \end{bmatrix} = \frac{1}{\sum_{j=1}^{K} \exp[w_j^T x]} \begin{bmatrix} \exp[w_1^T x] \\ \exp[w_2^T x] \\ \vdots \\ \exp[w_k^T x] \end{bmatrix}$$
\( X = \{ x^t, y^t \} \) \( y^t | x^t \sim \text{Multinomial}(...) \)

\[
o_k = \hat{P}(y=k|x) = \frac{\exp[w_k^T x]}{\sum_{j=1}^{K} \exp[w_j^T x]}, k = 1, \ldots, K
\]

Maximizing the likelihood is equivalent to minimizing the negative log likelihood (cross-entropy error)

\[
l(\{w_k\} | X) = \prod_t \prod_k (o_k^t)^{y_k^t}
\]

\[
E(\{w_k\} | X) = - \sum_{t=1}^{K} \sum_{k=1}^{K} 1\{y^t = k\} \log o_k^t = - \sum_{t=1}^{K} \sum_{k=1}^{K} 1\{y^t = k\} \log \frac{\exp[w_k^T x^t]}{\sum_{j=1}^{K} \exp[w_j^T x^t]}
\]

\( x \rightarrow O_k \) where \( k \) is the correct class