LOGISTIC REGRESSION

Consider a \textit{binary} classification problem with labels $y = 0, 1$.

Define
\[
\eta(x) = \frac{1}{1 + e^{-\langle w, x \rangle + b}}, \quad w \in \mathbb{R}^d, \quad b \in \mathbb{R}
\]

Then the Bayes classifier may be expressed as
\[
f^*(x) = \arg \max_{y \in \{0, 1\}} P(y|x)
\]

Logistic regression implements the following strategy:

1) Assume $\eta(x) = \frac{1}{1 + e^{-\langle w, x \rangle + b}}$, $w \in \mathbb{R}^d$, $b \in \mathbb{R}$

2) Compute the MLE of $\theta = (w, b)$.

3) Plug the estimate
\[
\hat{\eta}(x) = \frac{1}{1 + e^{-\langle \hat{w}, x \rangle + \hat{b}}}
\]

into the formula for the Bayes classifier.
The function $\frac{1}{1 + e^{-t}}$ is called a __________ function, and also called a ________ function.

Observe that

$$\hat{f}(x) = 1 \iff$$

Therefore

$$\hat{f}(x) =$$

is ________.
Maximum Likelihood Estimation

Assume the data \((x_i, y_i)\) are independent.

Denote \(x = (x_1, \ldots, x_m)\), \(y = (y_1, \ldots, y_n)\). Then

\[
\ell(\theta; x, y) =
\]

\[
= 
\]

\[
= 
\]
Note that $y$ is discrete and therefore 

$$(y \mid x; \theta)$$

is a probability mass function.

In particular, we recognize $y \mid x$ as a random variable with

$$p(y \mid x; \theta) = \begin{cases} \end{cases}$$

=
Maximum Likelihood Estimation

\[ l(\theta) = \prod_{i=1}^{n} \eta(x_i; \theta)^{y_i} (1 - \eta(x_i; \theta))^{1-y_i} + C \]

\[ \Rightarrow \log l(\theta) = \sum_{i=1}^{n} y_i \log \eta(x_i; \theta) + (1-y_i) \log (1-\eta(x_i; \theta)) \]

**Notation**

\[ \tilde{x} = [1 \ x^{(1)} \ ... \ x^{(d)}]^T \]

\[ \theta = [b \ w^{(1)} \ ... \ w^{(d)}]^T \]

\[ g(t) = \frac{1}{1+e^{-t}} \]

So that \[ \eta(x) = g(\theta^T \tilde{x}) \]

Note that \[ g'(t) = \]

So we have

\[ \log l(\theta) = \sum_{i} y_i \log g(\theta^T \tilde{x}_i) + (1-y_i) \log (1-g(\theta^T \tilde{x}_i)) \]
To maximize the likelihood, we can try

\[ \frac{\partial \log \ell(\theta)}{\partial \theta} = \sum_{i=1}^{n} \]

Unfortunately, this is a nonlinear system of equations and has no closed-form solution.

However, the log-likelihood is concave and therefore has a global maximum. Typically, the log-likelihood is maximized iteratively using the Newton–Raphson algorithm:

\[ \theta_{\text{new}} = \theta_{\text{old}} - \left( \frac{\partial^2 \log \ell(\theta)}{\partial \theta \partial \theta^T} \right)^{-1} \frac{\partial \log \ell(\theta)}{\partial \theta} \]

where derivatives are evaluated at \( \theta_{\text{old}} \)
A. \( \eta(x) = \Pr\{Y=1 \mid X=x\} \)
\[= 1 - \Pr\{Y=0 \mid X=x\} \]
\[f^*(x) = \begin{cases} 1 & \text{if } \eta(x) \geq \frac{1}{2} \\ 0 & \text{if } \eta(x) < \frac{1}{2} \end{cases} \]

B. logistic, sigmoid

C. \( \hat{f}(x) = 1 \iff \hat{\eta}(x) \geq \frac{1}{2} \)

\[\iff \exp\left\{- (\hat{w}^T x + \hat{b})\right\} \leq 1 \]

\[\iff \hat{w}^T x + \hat{b} \geq 0 \]

\[\hat{f}(x) = \begin{cases} 1 & \text{if } \hat{w}^T x + \hat{b} \geq 0 \\ 0 & \text{otherwise} \end{cases} \]

\(\Rightarrow \hat{f} \text{ is linear} \)

D. \( l(\theta; x, y) = p(x, y; \theta) \)
\[= \prod_{i=1}^{n} p(x_i, y_i; \theta) \]
\[= \prod_{i=1}^{n} p(y_i \mid x_i; \theta) \cdot p(x_i; \theta) \]
\[\text{independent of } \theta \]

E. \( g'(t) = \frac{e^{-t}}{(1 + e^{-t})^2} = g(t) \cdot (1 - g(t)) \)
\[
\frac{\partial \log \ell (\theta)}{\partial \theta} = \sum_{i=1}^{n} y_i \tilde{x}_i (1 - g(\theta^T \tilde{x}_i)) - (1-y_i) \tilde{x}_i g(\theta^T \tilde{x}_i)
\]
\[
= \sum_{i=1}^{n} \tilde{x}_i (y_i - g(\theta^T \tilde{x}_i))
\]
\[
= 0
\]