Logistic Regression, Exponential Family

Junxian He
Feb 9, 2024
Classification

Labels are discrete
Logistic Regression
Logistic Regression

Given a training set \( \{(x^{(i)}, y^{(i)}) \text{ for } i = 1, \ldots, n\} \) let \( y^{(i)} \in \{0, 1\} \). Want \( h_{\theta}(x) \in [0, 1] \). Let’s pick a smooth function:

\[
h_{\theta}(x) = g(\theta^T x)
\]
Logistic Regression

Given a training set \{ (x^{(i)}, y^{(i)}) \} for \( i = 1, \ldots, n \) let \( y^{(i)} \in \{0, 1\} \).
Want \( h_\theta(x) \in [0, 1] \). Let’s pick a smooth function:

\[
h_\theta(x) = g(\theta^T x) \quad \text{Link Function}
\]
Logistic Regression

Given a training set \( \{(x^{(i)}, y^{(i)}) \text{ for } i = 1, \ldots, n\} \) let \( y^{(i)} \in \{0, 1\} \). Want \( h_{\theta}(x) \in [0, 1] \). Let’s pick a smooth function:

\[
h_{\theta}(x) = g(\theta^T x) \quad \text{Link Function}
\]

There are many options of \( g \).
Logistic Regression

Given a training set \( \{(x^{(i)}, y^{(i)}) \text{ for } i = 1, \ldots, n\} \) let \( y^{(i)} \in \{0, 1\} \).

Want \( h_\theta(x) \in [0, 1] \). Let’s pick a smooth function:

\[
h_\theta(x) = g(\theta^T x) \quad \text{Link Function}
\]

There are many options of \( g \)....

\[
g(z) = \frac{1}{1 + e^{-z}}.
\]
Logistic Regression

Given a training set \( \{(x^{(i)}, y^{(i)}) \text{ for } i = 1, \ldots, n\} \) let \( y^{(i)} \in \{0, 1\} \). Want \( h_\theta(x) \in [0, 1] \). Let’s pick a smooth function:

\[
h_\theta(x) = g(\theta^T x) \quad \text{Link Function}
\]

There are many options of \( g \).

\[
g(z) = \frac{1}{1 + e^{-z}}.
\]
Logistic Regression

Given a training set \( \{(x^{(i)}, y^{(i)}) \} \) for \( i = 1, \ldots, n \) let \( y^{(i)} \in \{0, 1\} \). Want \( h_\theta(x) \in [0, 1] \). Let’s pick a smooth function:

\[
h_\theta(x) = g(\theta^T x) \quad \text{Link Function}
\]

There are many options of \( g \)....

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

How do we interpret \( h_\theta(x) \)?

\[
P(y = 1 \mid x; \theta) = h_\theta(x)
\]

\[
P(y = 0 \mid x; \theta) = 1 - h_\theta(x)
\]
Logistic Regression

Given a training set \( \{(x^{(i)}, y^{(i)}) \mid i = 1, \ldots, n\} \) let \( y^{(i)} \in \{0, 1\} \).
Want \( h_\theta(x) \in [0, 1] \). Let’s pick a smooth function:

\[
h_\theta(x) = g(\theta^T x)
\]

**Link Function**

There are many options of \( g \).

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

**Logistic Function**

How do we interpret \( h_\theta(x) \)?

\[
P(y = 1 \mid x; \theta) = h_\theta(x)
\]
\[
P(y = 0 \mid x; \theta) = 1 - h_\theta(x)
\]
Logistic Regression

Given a training set \( \{(x^{(i)}, y^{(i)}) \} \) for \( i = 1, \ldots, n \) let \( y^{(i)} \in \{0, 1\} \).
Want \( h_\theta(x) \in [0, 1] \). Let’s pick a smooth function:

\[
h_\theta(x) = g(\theta^T x) \quad \text{Link Function}
\]

There are many options of \( g \)....

\[
g(z) = \frac{1}{1 + e^{-z}} \cdot \quad \text{Logistic Function}
\]

\[
\text{Sigmoid Function}
\]

How do we interpret \( h_\theta(x) \)?

\[
P(y = 1 \mid x; \theta) = h_\theta(x)
\]

\[
P(y = 0 \mid x; \theta) = 1 - h_\theta(x)
\]
Logistic Regression

Let’s write the Likelihood function. Recall:

\[ P(y = 1 \mid x; \theta) = h_\theta(x) \]
\[ P(y = 0 \mid x; \theta) = 1 - h_\theta(x) \]

Maximum likelihood estimation
Logistic Regression

Let’s write the Likelihood function. Recall:

\[ P(y = 1 \mid x; \theta) = h_\theta(x) \]
\[ P(y = 0 \mid x; \theta) = 1 - h_\theta(x) \]

Then,

\[ L(\theta) = P(y \mid X; \theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta) \]

Maximum likelihood estimation
Logistic Regression

Let’s write the Likelihood function. Recall:

\[ P(y = 1 \mid x; \theta) = h_\theta(x) \]
\[ P(y = 0 \mid x; \theta) = 1 - h_\theta(x) \]

Then,

\[ L(\theta) = P(y \mid X; \theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta) \]

We want to express “if-then” logics, how?

Maximum likelihood estimation
Logistic Regression

Let’s write the Likelihood function. Recall:

\[
P(y = 1 \mid x; \theta) = h_\theta(x) \\
P(y = 0 \mid x; \theta) = 1 - h_\theta(x)
\]

Then,

\[
L(\theta) = P(y \mid X; \theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta) = \prod_{i=1}^{n} h_\theta(x^{(i)})^{y^{(i)}}(1 - h_\theta(x^{(i)}))^{1-y^{(i)}}
\]

We want to express “if-then” logics, how?

Maximum likelihood estimation
Logistic Regression

Let's write the Likelihood function. Recall:

\[ P(y = 1 \mid x; \theta) = h_\theta(x) \]
\[ P(y = 0 \mid x; \theta) = 1 - h_\theta(x) \]

Then,

\[ L(\theta) = P(y \mid X; \theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta) \]
\[ = \prod_{i=1}^{n} h_\theta(x^{(i)})^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}} \]

We want to express “if-then” logics, how?

Taking logs to compute the log likelihood \( \ell(\theta) \) we have:

\[ \ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \]

Maximum likelihood estimation
Derivative of Logistic Function

\[ g'(z) = \frac{d}{dz} \frac{1}{1 + e^{-z}} \]

\[ = \frac{1}{(1 + e^{-z})^2} (e^{-z}) \]

\[ = \frac{1}{(1 + e^{-z})} \cdot \left( 1 - \frac{1}{(1 + e^{-z})} \right) \]

\[ = g(z)(1 - g(z)). \]
Gradient Descent

\[
\frac{\partial}{\partial \theta_j} \ell(\theta) = \left( y \frac{1}{g(\theta^T x)} - (1 - y) \frac{1}{1 - g(\theta^T x)} \right) \frac{\partial}{\partial \theta_j} g(\theta^T x) \\
= \left( y \frac{1}{g(\theta^T x)} - (1 - y) \frac{1}{1 - g(\theta^T x)} \right) g(\theta^T x)(1 - g(\theta^T x)) \frac{\partial}{\partial \theta_j} \theta^T x \\
= (y(1 - g(\theta^T x)) - (1 - y)g(\theta^T x)) x_j \\
= (y - h_\theta(x)) x_j
\]

\[
\theta_j := \theta_j + \alpha \left( y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)}
\]
Gradient Descent

\[ \frac{\partial}{\partial \theta_j} \ell(\theta) = \left( y \frac{1}{g(\theta^T x)} - (1 - y) \frac{1}{1 - g(\theta^T x)} \right) \frac{\partial}{\partial \theta_j} g(\theta^T x) \]

\[ = \left( y \frac{1}{g(\theta^T x)} - (1 - y) \frac{1}{1 - g(\theta^T x)} \right) g(\theta^T x)(1 - g(\theta^T x)) \frac{\partial}{\partial \theta_j} \theta^T x \]

\[ = (y(1 - g(\theta^T x)) - (1 - y)g(\theta^T x)) x_j \]

\[ = (y - h_\theta(x)) x_j \]

\[ \theta_j := \theta_j + \alpha \left( y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)} \]

Looks identical to LMS update rule in linear regression
Gradient Descent

\[
\frac{\partial}{\partial \theta_j} \ell(\theta) = \left( y \frac{1}{g(\theta^T x)} - (1 - y) \frac{1}{1 - g(\theta^T x)} \right) \frac{\partial}{\partial \theta_j} g(\theta^T x)
\]

\[
= \left( y \frac{1}{g(\theta^T x)} - (1 - y) \frac{1}{1 - g(\theta^T x)} \right) g(\theta^T x)(1 - g(\theta^T x)) \frac{\partial}{\partial \theta_j} \theta^T x
\]

\[
= \left( y(1 - g(\theta^T x)) - (1 - y)g(\theta^T x) \right) x_j
\]

\[
= (y - h_\theta(x)) x_j
\]

\[
\theta_j := \theta_j + \alpha (y^{(i)} - h_\theta(x^{(i)})) x^{(i)}_j
\]

Looks identical to LMS update rule in linear regression

Is this coincidence?
Multi-Label Classification

{Cat, dog, dragon, fish, pig}
Multi-Label Classification

Given a training set \[\{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}\], \(y^{(i)} \in \{1, 2, \ldots, k\}\), we aim to model the distribution \(p(y \mid x; \theta)\).
Multi-Label Classification

Given a training set \( \{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\} \), \( y^{(i)} \in \{1,2,\ldots,k\} \), we aim to model the distribution \( p(y | x; \theta) \)

Categorical distribution, \( p(y = k | x; \theta) = \phi_k \)

\[
\text{s.t. } \sum_{i=1}^{k} \phi_i = 1
\]
Multi-Label Classification

Given a training set \( \{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\} \), \( y^{(i)} \in \{1, 2, \ldots, k\} \), we aim to model the distribution \( p(y \mid x; \theta) \).

Categorical distribution, \( p(y = k \mid x; \theta) = \phi_k \)

\[
\text{s.t. } \sum_{i=1}^{k} \phi_i = 1
\]

\[
\phi_i = \theta_i^T x ?
\]
Softmax Function
Softmax Function

Softmax: $\mathbb{R}^k \rightarrow \mathbb{R}^k$
So $\mathbb{R}^k \rightarrow \mathbb{R}^k$

$\text{softmax}(t_1, \ldots, t_k) = \begin{bmatrix}
\frac{\exp(t_1)}{\sum_{j=1}^k \exp(t_j)} \\
\vdots \\
\frac{\exp(t_k)}{\sum_{j=1}^k \exp(t_j)}
\end{bmatrix}$
Softmax Function

Softmax: $\mathbb{R}^k \rightarrow \mathbb{R}^k$

$$\text{softmax}(t_1, \ldots, t_k) = \left[ \frac{\exp(t_1)}{\sum_{j=1}^{k} \exp(t_j)} \right]$$

The denominator is a normalization constant.
Multi-Label Classification
Multi-Label Classification

Let \((t_1, \ldots, t_k) = (\theta_1^\top x, \cdots, \theta_k^\top x)\)
Multi-Label Classification

Let \((t_1, \ldots, t_k) = (\theta_1^\top x, \cdots, \theta_k^\top x)\)

\[
\begin{bmatrix}
P(y = 1 \mid x; \theta) \\
\vdots \\
P(y = k \mid x; \theta)
\end{bmatrix}
= \text{softmax}(t_1, \cdots, t_k) = 
\begin{bmatrix}
\frac{\exp(\theta_1^\top x)}{\sum_{j=1}^{k} \exp(\theta_j^\top x)} \\
\vdots \\
\frac{\exp(\theta_k^\top x)}{\sum_{j=1}^{k} \exp(\theta_j^\top x)}
\end{bmatrix}
\]
Multi-Label Classification

Let \((t_1, \ldots, t_k) = (\theta_1^\top x, \ldots, \theta_k^\top x)\)

\[
\begin{bmatrix}
P(y = 1 \mid x; \theta) \\
\vdots \\
P(y = k \mid x; \theta)
\end{bmatrix}
= \text{softmax}(t_1, \ldots, t_k) =
\begin{bmatrix}
\exp(\theta_1^\top x) \\
\sum_{j=1}^k \exp(\theta_j^\top x)
\end{bmatrix}
\begin{bmatrix}
\exp(\theta_2^\top x) \\
\sum_{j=1}^k \exp(\theta_j^\top x)
\end{bmatrix}
\end{equation}

\[
P(y = i \mid x; \theta) = \phi_i = \frac{\exp(t_i)}{\sum_{j=1}^k \exp(t_j)} = \frac{\exp(\theta_i^\top x)}{\sum_{j=1}^k \exp(\theta_j^\top x)}
\]
Multi-Label Classification
Multi-Label Classification

\[ -\log p(y \mid x, \theta) = -\log \left( \frac{\exp(t_y)}{\sum_{j=1}^{k} \exp(t_j)} \right) = -\log \left( \frac{\exp(\theta_y^T x)}{\sum_{j=1}^{k} \exp(\theta_j^T x)} \right) \]
Multi-Label Classification

\[- \log p(y \mid x, \theta) = - \log \left( \frac{\exp(t_y)}{\sum_{j=1}^{k} \exp(t_j)} \right) = - \log \left( \frac{\exp(\theta_y^\top x)}{\sum_{j=1}^{k} \exp(\theta_j^\top x)} \right)\]

\[\ell(\theta) = \sum_{i=1}^{n} - \log \left( \frac{\exp(\theta_{y(i)}^\top x^{(i)})}{\sum_{j=1}^{k} \exp(\theta_j^\top x^{(i)})} \right)\]
Multi-Label Classification

\[- \log p(y | x, \theta) = - \log \left( \frac{\exp(t_y)}{\sum_{j=1}^{k} \exp(t_j)} \right) = - \log \left( \frac{\exp(\theta_y^\top x)}{\sum_{j=1}^{k} \exp(\theta_j^\top x)} \right) \]

\[\ell(\theta) = \sum_{i=1}^{n} - \log \left( \frac{\exp(\theta_{y(i)}^\top x^{(i)})}{\sum_{j=1}^{k} \exp(\theta_j^\top x^{(i)})} \right) \]

Negative log likelihood
Multi-Label Classification

\[- \log p(y \mid x, \theta) = - \log \left( \frac{\exp(t_y)}{\sum_{j=1}^{k} \exp(t_j)} \right) = - \log \left( \frac{\exp(\theta_y^\top x)}{\sum_{j=1}^{k} \exp(\theta_j^\top x)} \right)\]

\[\ell(\theta) = \sum_{i=1}^{n} - \log \left( \frac{\exp(\theta_{y(i)}^\top x(i))}{\sum_{j=1}^{k} \exp(\theta_j^\top x(i))} \right)\]

Negative log likelihood

Cross-entropy loss \[\ell_{ce} : \mathbb{R}^k \times \{1, \ldots, k\} \rightarrow \mathbb{R}_{\geq 0}\]
### Multi-Label Classification

\[- \log p(y \mid x, \theta) = - \log \left( \frac{\exp(t_y)}{\sum_{j=1}^{k} \exp(t_j)} \right) = - \log \left( \frac{\exp(\theta_y^\top x)}{\sum_{j=1}^{k} \exp(\theta_j^\top x)} \right)\]

\[
\ell(\theta) = \sum_{i=1}^{n} - \log \left( \frac{\exp(\theta_{y(i)}^\top x^{(i)})}{\sum_{j=1}^{k} \exp(\theta_j^\top x^{(i)})} \right)
\]

Negative log likelihood

Cross-entropy loss

\[\ell_{ce} : \mathbb{R}^k \times \{1, \ldots, k\} \rightarrow \mathbb{R}_{\geq 0}\]

\[\ell_{ce}((t_1, \ldots, t_k), y) = - \log \left( \frac{\exp(t_y)}{\sum_{j=1}^{k} \exp(t_j)} \right)\]
Multi-Label Classification

\[- \log p(y \mid x, \theta) = - \log \left( \frac{\exp(t_y)}{\sum_{j=1}^k \exp(t_j)} \right) = - \log \left( \frac{\exp(\theta_y^\top x)}{\sum_{j=1}^k \exp(\theta_j^\top x)} \right)\]

\[\ell(\theta) = \sum_{i=1}^n - \log \left( \frac{\exp(\theta_{y(i)}^\top x^{(i)})}{\sum_{j=1}^k \exp(\theta_j^\top x^{(i)})} \right)\]

Negative log likelihood

Cross-entropy loss \( \ell_{\text{ce}} : \mathbb{R}^k \times \{1, \ldots, k\} \to \mathbb{R}_{\geq 0} \)

\[\ell_{\text{ce}}((t_1, \ldots, t_k), y) = - \log \left( \frac{\exp(t_y)}{\sum_{j=1}^k \exp(t_j)} \right)\]

\[\ell(\theta) = \sum_{i=1}^n \ell_{\text{ce}}((\theta_1^\top x^{(i)}, \ldots, \theta_k^\top x^{(i)}), y^{(i)})\]
The Derivative
The Derivative

\[ \frac{\partial \ell_{ce}(t, y)}{\partial t_i} = \phi_i - 1\{y = i\} \]
The Derivative

\[ \frac{\partial \ell_{ce}(t, y)}{\partial t_i} = \phi_i - 1 \{ y = i \} \]

\[ \phi_i = \frac{\exp(t_i)}{\sum_{j=1}^{k} \exp(t_j)} \]
The Derivative

\[
\frac{\partial \ell_{ce}(t, y)}{\partial t_i} = \phi_i - 1\{y = i\}
\]

\[
\phi_i = \frac{\exp(t_i)}{\sum_{j=1}^{k} \exp(t_j)}
\]

Chain rule

\[
\frac{\partial \ell_{ce}((\theta_1^t x, \ldots, \theta_k^t x), y)}{\partial \theta_i} = \frac{\partial \ell(t, y)}{\partial t_i} \cdot \frac{\partial t_i}{\partial \theta_i} = (\phi_i - 1\{y = i\}) \cdot x
\]
The Derivative

\[ \frac{\partial \ell_{ce}(t, y)}{\partial t_i} = \phi_i - 1\{y = i\} \quad \quad \quad \phi_i = \frac{\exp(t_i)}{\sum_{j=1}^{k} \exp(t_j)} \]

Chain rule

\[ \frac{\partial \ell_{ce}(\theta_1^T x, \ldots, \theta_k^T x), y)}{\partial \theta_i} = \frac{\partial \ell(t, y)}{\partial t_i} \cdot \frac{\partial t_i}{\partial \theta_i} = (\phi_i - 1\{y = i\}) \cdot x \]

\[ \frac{\partial \ell(\theta)}{\partial \theta_i} = \sum_{j=1}^{n} (\phi_i^{(j)} - 1\{y^{(j)} = i\}) \cdot x^{(j)} \]
The Derivative

\[
\frac{\partial \ell_{ce}(t, y)}{\partial t_i} = \phi_i - 1\{y = i\}
\]

\[
\phi_i = \frac{\exp(t_i)}{\sum_{j=1}^{k} \exp(t_j)}
\]

Chain rule

\[
\frac{\partial \ell_{ce}((\theta_1^T x, \ldots, \theta_k^T x), y)}{\partial \theta_i} = \frac{\partial \ell(t, y)}{\partial t_i} \cdot \frac{\partial t_i}{\partial \theta_i} = (\phi_i - 1\{y = i\}) \cdot x
\]

\[
\frac{\partial \ell(\theta)}{\partial \theta_i} = \sum_{j=1}^{n} (\phi_i^{(j)} - 1\{y^{(j)} = i\}) \cdot x^{(j)}
\]

Intuitive explanation of the rule?
Another Optimization Method — Newton’s Method
Another Optimization Method — Newton’s Method

Given $f : \mathbb{R}^d \to \mathbb{R}$ find $x$ s.t. $f(x) = 0$. 
Another Optimization Method — Newton’s Method

Given \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) find \( x \) s.t. \( f(x) = 0 \). \[ \nabla_\theta l(\theta) = 0 \]
Another Optimization Method — Newton’s Method

Given $f : \mathbb{R}^d \rightarrow \mathbb{R}$ find $x$ s.t. $f(x) = 0$.  
\[ \nabla_\theta l(\theta) = 0 \]

- This is the update rule in 1d

\[
    x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})}
\]
Another Optimization Method — Newton’s Method

Given \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) find \( x \) s.t. \( f(x) = 0 \).

\[ \nabla_\theta l(\theta) = 0 \]

This is the update rule in 1d

\[ x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})} \]

Solution to a linear equation

\[ f'(x^{(t)})x^{(t+1)} + f(x^{(t)}) - x^{(t)}f'(x^{(t)}) = 0 \]
Another Optimization Method — Newton’s Method

Given $f : \mathbb{R}^d \to \mathbb{R}$ find $x$ s.t. $f(x) = 0$.  

\[ \nabla_\theta l(\theta) = 0 \]

▶ This is the update rule in 1d

\[ x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})} \]

Solution to a linear equation

\[ f'(x^{(t)})x^{(t+1)} + f(x^{(t)}) - x^{(t)}f'(x^{(t)}) = 0 \]

View it as an equation of $x^{(t+1)}$, and $x^{(t)}$ is a constant
Another Optimization Method — Newton’s Method

\[ f'(x^{(t)})x + f(x^{(t)}) - x^{(t)}f'(x^{(t)}) = y \]

\[(x^{(t)}, f(x^{(t)}))\]
Another Optimization Method — Newton’s Method

\[ f'(x^{(t)})x + f(x^{(t)}) - x^{(t)}f'(x^{(t)}) = y \]

\[ (x^{(t)}, f(x^{(t)})) \]
Another Optimization Method — Newton’s Method

\[ f'(x^{(t)})x + f(x^{(t)}) - x^{(t)}f'(x^{(t)}) = y \]

\((x^{(t)}, f(x^{(t)}))\)
Another Optimization Method — Newton’s Method

\[ f'(x^{(t)})x + f(x^{(t)}) - x^{(t)}f'(x^{(t)}) = y \]

\[ (x^{(t)}, f(x^{(t)})) \]
Another Optimization Method — Newton’s Method

Given $f : \mathbb{R}^d \rightarrow \mathbb{R}$ find $x$ s.t. $f(x) = 0$.  
$\nabla_\theta l(\theta) = 0$

- This is the update rule in 1d

$$x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})}$$

$$\theta := \theta - \frac{\ell'(\theta)}{\ell''(\theta)}.$$
Another Optimization Method — Newton’s Method

Given \( f : \mathbb{R}^d \to \mathbb{R} \) find \( x \) s.t. \( f(x) = 0 \). \( \nabla_\theta l(\theta) = 0 \)

- This is the update rule in 1d

\[
    x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})}
\]

- It may converge *very* fast (quadratic local convergence!) Requires fewer iterations
Another Optimization Method — Newton’s Method

Given $f : \mathbb{R}^d \rightarrow \mathbb{R}$ find $x$ s.t. $f(x) = 0$. \[ \nabla_\theta l(\theta) = 0 \]

- This is the update rule in 1d

\[ x(t+1) = x(t) - \frac{f(x(t))}{f'(x(t))} \]

\[ \theta := \theta - \frac{\ell'(\theta)}{\ell''(\theta)}. \]

- It may converge very fast (quadratic local convergence!) Requires fewer iterations

- For the likelihood, i.e., $f(\theta) = \nabla_\theta \ell(\theta)$ we need to generalize to a vector-valued function which has:

\[ \theta^{(t+1)} = \theta^{(t)} - \left( H(\theta^{(t)}) \right)^{-1} \nabla_\theta \ell(\theta^{(t)}). \]

in which $H_{i,j}(\theta) = \frac{\partial}{\partial \theta_i \partial \theta_j} \ell(\theta)$. 

15
Another Optimization Method — Newton’s Method

Given $f : \mathbb{R}^d \rightarrow \mathbb{R}$ find $x$ s.t. $f(x) = 0$. \[ \nabla_\theta l(\theta) = 0 \]

- This is the update rule in 1d

\[ x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})} \]

\[ \theta := \theta - \frac{\ell'(\theta)}{\ell''(\theta)}. \]

- It may converge very fast (quadratic local convergence!) Requires fewer iterations

- For the likelihood, i.e., $f(\theta) = \nabla_\theta \ell(\theta)$ we need to generalize to a vector-valued function which has:

\[ \theta^{(t+1)} = \theta^{(t)} - \left( H(\theta^{(t)}) \right)^{-1} \nabla_\theta \ell(\theta^{(t)}). \]

in which $H_{i,j}(\theta) = \frac{\partial}{\partial \theta_i \partial \theta_j} \ell(\theta)$. When Newton’s method is applied to maximize the logistic regression log likelihood function $l(\theta)$, the resulting method is also called Fisher scoring.
Exponential Family
Exponential Family

- Exponential family unifies inference and learning for many important models
Exponential Family
Exponential Family

Rough Idea "If $P$ has a special form, then inference and learning come for free"

$$P(y; \eta) = b(y) \exp \left\{ \eta^T T(y) - a(\eta) \right\}.$$ 

Here $y$, $a(\eta)$, and $b(y)$ are scalars. $T(y)$ same dimension as $\eta$. 
Exponential Family

**Rough Idea** "If $P$ has a a special form, then inference and learning come for free"

$$P(y; \eta) = b(y) \exp \left\{ \eta^T T(y) - a(\eta) \right\}.$$

*\eta*: natural parameter or canonical parameter

Here $y$, $a(\eta)$, and $b(y)$ are scalars. $T(y)$ same dimension as $\eta$. 
Exponential Family

Rough Idea “If $P$ has a a special form, then inference and learning come for free”

\[ P(y; \eta) = b(y) \exp \left\{ \eta^T T(y) - a(\eta) \right\}. \]

$\eta$: natural parameter or canonical parameter

Here $y$, $a(\eta)$, and $b(y)$ are scalars. $T(y)$ same dimension as $\eta$.

$T(y)$ is called the sufficient statistic.

$b(y)$ is called the base measure – does not depend on $\eta$.

$a(\eta)$ is called the log partition function – does not depend on $y$. 
Exponential Family

Rough Idea “If $P$ has a special form, then inference and learning come for free”

$$P(y; \eta) = b(y) \exp \left\{ \eta^T T(y) - a(\eta) \right\}.$$  
$\eta$: natural parameter or canonical parameter

Here $y$, $a(\eta)$, and $b(y)$ are scalars. $T(y)$ same dimension as $\eta$.

$T(y)$ is called the **sufficient statistic**.  
holds all information the data provides with regard to the unknown parameter values  

$b(y)$ is called the **base measure** – does not depend on $\eta$.

$a(\eta)$ is called the **log partition function** – does not depend on $y$.  

Exponential Family

**Rough Idea** “If $P$ has a special form, then inference and learning come for free”

$$ P(y; \eta) = b(y) \exp \left\{ \eta^T T(y) - a(\eta) \right\}. $$

$\eta$: natural parameter or canonical parameter

Here $y$, $a(\eta)$, and $b(y)$ are scalars. $T(y)$ same dimension as $\eta$.

$T(y)$ is called the **sufficient statistic**. $b(y)$ is called the **base measure** – does **not** depend on $\eta$. $a(\eta)$ is called the **log partition function** – does **not** depend on $y$.

$$ 1 = \sum_y P(y; \eta) = e^{-a(\eta)} \sum_y b(y) \exp \left\{ \eta^T T(y) \right\} $$

$$ \implies a(\eta) = \log \sum_y b(y) \exp \left\{ \eta^T T(y) \right\} $$
Example: Bernoulli

Bernoulli random variable is an event (say flipping a coin) then:

\[ p(y; \phi) = \phi^y (1 - \phi)^{1-y} \]
Example: Bernoulli

Bernoulli random variable is an event (say flipping a coin) then:

\[ p(y; \phi) = \phi^y (1 - \phi)^{1-y} \]

How do we put it in the required form?

\[ P(y; \eta) = b(y) \exp \left\{ \eta^T T(y) - a(\eta) \right\} . \]
Example: Bernoulli

Bernoulli random variable is an event (say flipping a coin) then:

$$p(y; \phi) = \phi^y (1 - \phi)^{1-y}$$

How do we put it in the required form?

$$P(y; \eta) = b(y) \exp \left\{ \eta^T T(y) - a(\eta) \right\}.$$  

$$p(y; \phi) = \phi^y (1 - \phi)^{1-y}$$

$$= \exp(y \log \phi + (1 - y) \log(1 - \phi))$$

$$= \exp \left( \left( \log \left( \frac{\phi}{1 - \phi} \right) \right) y + \log(1 - \phi) \right)$$
Example: Bernoulli

\[ P(y; \eta) = b(y) \exp \left\{ \eta^T T(y) - a(\eta) \right\} \]

\[
p(y; \phi) = \phi^y (1 - \phi)^{1-y}
\]

\[
= \exp (y \log \phi + (1 - y) \log (1 - \phi))
\]

\[
= \exp \left( \left( \log \left( \frac{\phi}{1 - \phi} \right) \right) y + \log (1 - \phi) \right)
\]
Example: Bernoulli

\[ P(y; \eta) = b(y) \exp \left\{ \eta^T T(y) - a(\eta) \right\} \]

\[
\begin{align*}
p(y; \phi) & = \phi^y (1 - \phi)^{1-y} \\
& = \exp(y \log \phi + (1 - y) \log(1 - \phi)) \\
& = \exp \left( \left( \log \left( \frac{\phi}{1 - \phi} \right) \right) y + \log(1 - \phi) \right)
\end{align*}
\]

So then:

\[
\eta = \log \frac{\phi}{1 - \phi}, \quad T(y) = y, \quad a(\eta) = -\log(1 - \phi).
\]

\[ b(y) = 1 \]
Example: Bernoulli

\[ P(y; \eta) = b(y) \exp \left\{ \eta^T T(y) - a(\eta) \right\} \]

\[
\begin{align*}
p(y; \phi) & = \phi^y (1 - \phi)^{1-y} \\
& = \exp(y \log \phi + (1 - y) \log(1 - \phi)) \\
& = \exp \left( \left( \log \left( \frac{\phi}{1 - \phi} \right) \right) y + \log(1 - \phi) \right)
\end{align*}
\]

So then:

\[
\eta = \log \frac{\phi}{1 - \phi}, \quad T(y) = y, \quad a(\eta) = -\log(1 - \phi).
\]

\[ b(y) = 1 \]

We need to show \( a(\eta) \) is a function of \( \log \frac{\phi}{1 - \phi} \)
Example: Bernoulli
Example: Bernoulli

We first observe that:

\[ \eta = \log \frac{\phi}{1 - \phi} \implies e^\eta (1 - \phi) = \phi \]

\[ e^\eta = (e^\eta + 1)\phi \implies \phi = \frac{1}{1 + e^{-\eta}} \]
Example: Bernoulli

We first observe that:

\[ \eta = \log \frac{\phi}{1 - \phi} \implies e^{\eta}(1 - \phi) = \phi \]

\[ e^{\eta} = (e^{\eta} + 1)\phi \implies \phi = \frac{1}{1 + e^{-\eta}} \]

Now, we plug into \( \log(1 - \phi) \) and we verify:

\[ a(\eta) = \log(1 - \phi) = \log \frac{e^{-\eta}}{1 + e^{-\eta}} = -\log(1 + e^{\eta}). \]
Example: Bernoulli

We first observe that:

\[ \eta = \log \frac{\phi}{1 - \phi} \implies e^\eta (1 - \phi) = \phi \]
\[ e^\eta = (e^\eta + 1)\phi \implies \phi = \frac{1}{1 + e^{-\eta}} \]

Now, we plug into \( \log(1 - \phi) \) and we verify:

\[ a(\eta) = \log(1 - \phi) = \log \frac{e^{-\eta}}{1 + e^{-\eta}} = -\log(1 + e^\eta) \]

We have verified Bernoulli distribution is in the exponential family.
Example: Gaussian with Fixed Variance $\sigma^2 = 1$
Example: Gaussian with Fixed Variance $\sigma^2 = 1$

$$P(y; \mu) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} (y - \mu)^2 \right\}.$$  

Can we put it in the exponential family form?

$$P(y; \eta) = b(y) \exp \left\{ \eta^T T(y) - a(\eta) \right\}.$$
Example: Gaussian with Fixed Variance $\sigma^2 = 1$

$$P(y; \mu) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} (y - \mu)^2 \right\}.$$  

Can we put it in the exponential family form?

$$P(y; \eta) = b(y) \exp \left\{ \eta^T T(y) - a(\eta) \right\}.$$ 

Multiply out the square and group terms:

$$P(y; \mu) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{y^2}{2} \right\} \exp \left\{ \mu y - \frac{1}{2} \mu^2 \right\}.$$
Example: Gaussian with Fixed Variance $\sigma^2 = 1$

$$P(y; \mu) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} (y - \mu)^2 \right\}.$$  

Can we put it in the exponential family form?

$$P(y; \eta) = b(y) \exp \left\{ \eta^T T(y) - a(\eta) \right\}.$$  

Multiply out the square and group terms:

$$P(y; \mu) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -y^2 / 2 \right\} \exp \left\{ \mu y - \frac{1}{2} \mu^2 \right\}.$$  

$$\eta = \mu, \ T(y) = y, \ a(\eta) = \frac{1}{2} \eta^2.$$
An Observation
An Observation

Notice that for a Gaussian with mean $\mu$ we had

$$\eta = \mu, \ T(y) = y, \ a(\eta) = \frac{1}{2}\eta^2.$$
An Observation

Notice that for a Gaussian with mean $\mu$ we had

$$\eta = \mu, \ T(y) = y, \ a(\eta) = \frac{1}{2} \eta^2.$$  

$$\partial_\eta a(\eta) = \eta = \mu = \mathbb{E}[y] \text{ and } \partial^2_\eta a(\eta) = 1 = \sigma^2 = \text{var}(y)$$
An Observation

Notice that for a Gaussian with mean $\mu$ we had

$$\eta = \mu, \ T(y) = y, \ a(\eta) = \frac{1}{2}\eta^2.$$ 

$$\partial_\eta a(\eta) = \eta = \mu = \mathbb{E}[y] \text{ and } \partial^2_\eta a(\eta) = 1 = \sigma^2 = \text{var}(y)$$

Is this true for general?
Log Partition Function

Yes! Recall that

\[ a(\eta) = \log \sum_{y} b(y) \exp \left\{ \eta^T T(y) \right\} \]
Log Partition Function

Yes! Recall that

$$a(\eta) = \log \sum_y b(y) \exp \left\{ \eta^T T(y) \right\}$$

Then, taking derivatives

$$\nabla_{\eta} a(\eta) = \frac{\sum_y T(y) b(y) \exp \left\{ \eta^T T(y) \right\}}{\sum_y b(y) \exp \left\{ \eta^T T(y) \right\}} = \mathbb{E}[T(y); \eta]$$
Many Other Exponential Models

- There are many canonical exponential family models:
  - Binary $\leftrightarrow$ Bernoulli
  - Multiple Classes $\leftrightarrow$ Multinomial
  - Real $\leftrightarrow$ Gaussian
  - Counts $\leftrightarrow$ Poisson
  - $\mathbb{R}_+$ $\leftrightarrow$ Gamma, Exponential
  - Distributions $\leftrightarrow$ Dirichlet
Thank You!

Q & A