GLMs and Logistic Regression

Jonathan Pillow

Mathematical Tools for Neuroscience (NEU 314)
Fall, 2021

lecture 21
warm-up problems

Regression
1. Write down the formula for the least-squares regression solution for weights \( w \) given a design matrix \( X \) and output vector \( Y \)

KL divergence
2. Write down the formula for \( KL(P,Q) \), KL divergence between \( P \) & \( Q \).
3. When is \( KL(P,Q) \) zero?
4. When is \( KL(P,Q) \) infinite?
5. Compute the \( KL(P,Q) \) for distributions:
   \[
   P = [0.5, 0.5, 0] \\
   Q = [0.25, 0.25, 0.5]
   \]
6. Can you describe what this means in terms of yes/no questions?
The **exponential distribution** describes interspike intervals in a Poisson process (which is famously “memoryless”, meaning that how long you’ve been waiting provides no information about the next spike time).

\[
P(t \mid a) = ae^{-at}
\]

**Problem:** Compute the maximum likelihood estimator for the parameter ‘\(a\)’ of an given a set of \(N\) observed interspike intervals: \(\{t_1, t_2, \ldots, t_n\}\).
Example 1: linear Poisson neuron

spike count \( y \sim \text{Poisson}(\lambda) \)

spike rate \( \lambda = \theta x \)

parameter stimulus

encoding model:

\[
P(y|x, \theta) = \frac{1}{y!} \lambda^y e^{-\lambda} = \frac{1}{y!} (\theta x)^y e^{-(\theta x)}
\]
mean\( (y) = \theta x \)
var\( (y) = \theta x \)

conditional distribution

\( p(y|x = 35) \)
log \( P(Y|X, \theta) = \sum_i \log P(y_i|x_i, \theta) \)

\[ = \sum y_i \log \theta - \theta x_i + c \]

\[ = \log \theta (\sum y_i) - \theta (\sum x_i) \]

- Closed-form solution:

\[ \frac{d}{d\theta} \log P(Y|X, \theta) = \frac{1}{\theta} \sum y_i - \sum x_i = 0 \]

\[ \implies \hat{\theta}_{ML} = \frac{\sum y_i}{\sum x_i} \]

(let's notice: this is kind of a weird result!)
Example 2: linear Gaussian neuron

- **spike count**  \[ y \sim \mathcal{N}(\mu, \sigma^2) \]
- **spike rate**  \[ \mu = \theta x \]
  - parameter
  - stimulus

**Encoding model:**
\[
P(y|x, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\theta x)^2}{2\sigma^2}}
\]
\[
\text{mean}(y) = \theta x \\
\text{var}(y) = \sigma^2
\]

All slices have same width

encoding distribution

\[p(y|x = 20)\]
\[ P(y|x, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\theta x)^2}{2\sigma^2}} \]

Log-Likelihood
\[ \log P(Y|X, \theta) = - \sum \frac{(y_i - \theta x_i)^2}{2\sigma^2} + c \]

Do it: differentiate, set to zero, and solve for \( \theta \).
Log-Likelihood

\[
\log P(Y \mid X, \theta) = -\sum \frac{(y_i - \theta x_i)^2}{2\sigma^2} + c
\]

\[
\frac{d}{d\theta} \log P(Y \mid X, \theta) = -\sum \frac{(y_i - \theta x_i)x_i}{\sigma^2} = 0
\]

\[
\sum y_i x_i - \sum \theta x_i^2 = 0
\]

\[
\theta \sum x_i^2 = \sum y_i x_i
\]

Maximum-Likelihood Estimator:

(“Least squares regression” solution)

\[
\hat{\theta}_{ML} = \frac{\sum y_i x_i}{\sum x_i^2}
\]

(Recall that for Poisson, \(\hat{\theta}_{ML} = \frac{\sum y_i}{\sum x_i}\))
Log-Likelihood

\[ P(y|x, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\theta x)^2}{2\sigma^2}} \]

Log-Likelihood

\[ \log P(Y|X, \theta) = -\sum \frac{(y_i - \theta x_i)^2}{2\sigma^2} + c \]

\[ \frac{d}{d\theta} \log P(Y|X, \theta) = -\sum \frac{(y_i - \theta x_i)x_i}{\sigma^2} = 0 \]

\[ \sum y_i x_i - \sum \theta x_i^2 = 0 \]

\[ \theta \sum x_i^2 = \sum y_i x_i \]

Maximum-Likelihood Estimator:

(“Least squares regression” solution)

\[ \hat{\theta}_{ML} = \frac{\sum y_i x_i}{\sum x_i^2} \]

Matrix version:

\[ \hat{\theta}_{ML} = (X^T X)^{-1} X^T Y \]

(this is just least-squares regression!)
least-squares revisited

(switching $\theta$ to $\vec{k}$)

model:

$$y_t = \vec{k} \cdot \vec{x}_t + \epsilon_t$$

$\mathcal{N}(0, \sigma^2)$

Gaussian noise with variance $\sigma^2$

$$Y = X\vec{k} + \text{noise}$$

$\begin{bmatrix}
0 \\
0 \\
1 \\
\vdots
\end{bmatrix}
= \begin{bmatrix}
\vdots \\
\vdots \\
\end{bmatrix}
\begin{bmatrix}
\vec{k}
\end{bmatrix} + \text{noise}$

design matrix
least-squares revisited

model: \[ y_t = \mathbf{k} \cdot \mathbf{x}_t + \epsilon_t \]

equivalent to writing: \[ y_t | \mathbf{x}_t, \mathbf{k} \sim \mathcal{N}(\mathbf{x}_t \cdot \mathbf{k}, \sigma^2) \]

or

\[ p(y_t | \mathbf{x}_t, \mathbf{k}) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(y_t - \mathbf{x}_t \cdot \mathbf{k})^2}{2\sigma^2}} \]

For entire dataset: \[ p(Y | X, \mathbf{k}) = \prod_{t=1}^{T} p(y_t | \mathbf{x}_t, \mathbf{k}) \]

\[ = (2\pi\sigma^2)^{-\frac{T}{2}} \exp\left(-\sum_{t=1}^{T} \frac{(y_t - \mathbf{x}_t \cdot \mathbf{k})^2}{2\sigma^2}\right) \]

\[ \log P(Y | X, \mathbf{k}) = - \sum_{t=1}^{T} \frac{(y_t - \mathbf{x}_t \cdot \mathbf{k})^2}{2\sigma^2} + \text{const} \]

Guassian noise with variance \( \sigma^2 \)
least-squares revisited

model: \[ y_t = \bar{k} \cdot \bar{x}_t + \epsilon_t \]

equivalent to writing:
\[ y_t | \bar{x}_t, \bar{k} \sim \mathcal{N}(\bar{x}_t \cdot \bar{k}, \sigma^2) \]

or
\[
p(y_t | \bar{x}_t, \bar{k}) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(y_t - \bar{x}_t \cdot \bar{k})^2}{2\sigma^2}}
\]

**General points:**
1. minimizing a sum of squares is *always* equivalent to maximizing likelihood under a Gaussian noise model!
2. solution doesn’t depend on the noise variance \( \sigma^2 \)

\[ \log P(Y | X, \bar{k}) = - \sum_{t=1}^{T} \frac{(y_t - \bar{x}_t \cdot \bar{k})^2}{2\sigma^2} + \text{const} \]

\[ \mathcal{N}(0, \sigma^2) \]

Guassian noise with variance \( \sigma^2 \)
Example 3: unknown neuron

Be the computational neuroscientist: what model would you use?
Example 3: unknown neuron

More general setup: \( y \sim \text{Poisson}(\lambda) \)

\[ \lambda = f(\theta x), \] for some nonlinear function \( f \)
Example 3: unknown neuron

More general setup:

\[ y \sim \text{Poisson}(\lambda) \]

\[ \lambda = f(\theta x) \], for some nonlinear function \( f \)
Note on GLMs

• Be careful about terminology:

GLM ≠ GLM

General Linear Model ≠ Generalized Linear Model

(Nelder 1972)
1. General Linear Model

Examples:
1. Gaussian

\[ y = \theta \cdot \bar{x} + \epsilon \]

2. Poisson

\[ y \sim \text{Poiss}(\theta \cdot \bar{x}) \]
2. Generalized Linear Model

Examples:
1. Gaussian
   \[ y = f(\theta \cdot \vec{x}) + \epsilon \]
2. Poisson
   \[ y \sim \text{Poisson}(f(\theta \cdot \vec{x})) \]
aside:
Regression vs Classification
Classification

- mapping from vector input to discrete category

\[ f(\vec{x}) \]

\[ y = 0 \]

\[ y = 1 \]

\[ \vec{x} \cdot \vec{w} - b > 0 \]

linear classifier

- linear perceptron
- Fisher linear discriminant
- support vector machine (SVM)

(voxel activity)
(spike counts)
Regression

- output continuous instead of discrete

\[ y = \mathbf{x} \cdot \mathbf{w} \]

• can transform classification problems into regression problems (“logistic regression”):

\[ p(y = 1) = f(\mathbf{x}) \]

**Input**

\[ \mathbf{x} \]

(neural activity)

**Decision**

\[ f(\mathbf{x}) \rightarrow y \]

(hand position)

**Voxel 1**

**Voxel 2**

Monday, October 6, 14
GLM for binary responses

Bernoulli GLM:
(coin flipping model)

\[ p(y_t = 1|\vec{x}_t) = p_t \]
\[ p(y_t = 0|\vec{x}_t) = 1 - p_t \]
\[ p(y_t|\vec{x}_t) = p_t^{y_t}(1-p_t)^{1-y_t} \]
GLM for binary responses

**Bernoulli GLM:**

(coin flipping model)

Equivalent ways of writing:

\[
p(y_t = 1|\vec{x}_t) = p_t
\]

\[
p(y_t = 0|\vec{x}_t) = 1 - p_t
\]

\[
p(y_t|\vec{x}_t) = p_t^{y_t} (1 - p_t)^{1-y_t}
\]

Equivalent ways of writing:

\[
y_t|\vec{x}_t, \vec{k} \sim \text{Ber}(f(\vec{x}_t \cdot \vec{k}))
\]

or

\[
p(y_t|\vec{x}_t, \vec{k}) = f(\vec{x}_t \cdot \vec{k})^{y_t} \left(1 - f(\vec{x}_t \cdot \vec{k})\right)^{1-y_t}
\]

log-likelihood:

\[
\mathcal{L} = \sum_{t=1}^{T} \left( y_t \log f(\vec{x}_t \cdot \vec{k}) + (1 - y_t) \log(1 - f(\vec{x}_t \cdot \vec{k})) \right)
\]

In python:

\[
L = \text{np.sum( } Y*\text{np.log(f(X@k)) } + (1-Y)*\text{np.log(1-f(X@k))) }
\]
Bernoulli GLM: (coin flipping model)

\[
p(y_t = 1|x_t) = p_t
\]

\[
p(y_t = 0|x_t) = 1 - p_t
\]

\[
p(y_t | x_t) = p_t^{y_t} (1 - p_t)^{1-y_t}
\]

Logistic regression:

\[
f(x) = \frac{1}{1 + e^{-x}}
\]

- so logistic regression is a special case of a Bernoulli GLM, where the nonlinearity \( f(x) \) is a logistic function!
Summary (last 3 lectures)

• Estimation
• Bias
• Variance
• Maximum Likelihood estimator
• MAP estimation: accounts for slow-speed bias in motion perception (Weiss, Simoncelli & Adelson 2002)
• General and Generalized Linear Models (GLMs)
• Bernoulli GLM / Logistic regression