Recap

• (Ch 9) Inferring a probability model from a dataset
  • Maximum likelihood estimation (MLE)
  • Confidence intervals for MLE estimates

Today

• (Ch 9) Inferring a probability model from a dataset
  • Bayesian inference
  • Conjugate priors
• Review of eigenvalues, eigenvectors and diagonalization
Maximum likelihood estimation (MLE)

• We write the probability of seeing the data \( D \) given parameters \( \theta \)

\[
L(\theta) = P(D|\theta)
\]

• The **likelihood function** \( L(\theta) \) is not a probability distribution

• The **maximum likelihood estimate** of \( \theta \) is

\[
\hat{\theta} = \arg \max_{\theta} L(\theta)
\]
MLE: binomial example

- Suppose we have a coin of unknown probability $\theta$ of heads
  \[ \mathcal{N}_k \]
- We toss it 10 times and observe 7 heads
- The likelihood function is
  \[ L(\theta) = P(D|\theta) = \binom{10}{7} \theta^7 (1 - \theta)^3 \]
- The MLE is $\hat{\theta} = 0.7$
Drawbacks of MLE

• Maximizing some likelihood or log-likelihood functions is intractable

• If there isn’t much data, the MLE estimate may be unreliable
  • If we observe 3 heads in 10 coin tosses, should we accept that $P(\text{heads}) = 0.3$?
  • If we observe 0 heads in 2 coin tosses, should we accept that $P(\text{heads}) = 0$?
Bayesian inference

• In MLE, we maximized the likelihood function \( L(\theta) = P(D|\theta) \)

• In Bayesian inference, we will maximize the posterior, which is the probability of the parameters \( \theta \) given the observed data \( D \)

\[ P(\theta|D) \]

• Unlike \( L(\theta) \), the posterior is a probability distribution

• The value of \( \theta \) that maximizes \( P(\theta|D) \) is called the maximum a posteriori (MAP) estimate \( \hat{\theta} \)
The prior

• From Bayes rule

\[ P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \propto P(D|\theta)P(\theta) \]

• We ignore the probability of the data \( P(D) \) because it is constant

• Bayesian inference allows us to incorporate prior beliefs about \( \theta \) in the prior \( P(\theta) \), which is useful
  • when we have some beliefs, such as a coin cannot have \( P(\text{heads}) = 0 \)
  • when there isn’t much data
Bayesian inference: discrete prior example

- Suppose we have a coin of unknown probability $\theta$ of heads
  - We see heads 7 times in 10 tosses as the data $D$
  - Say we also have prior information about $\theta$: $P(\theta) = \begin{cases} \frac{2}{3} & \text{if } \theta = 0.5 \\ \frac{1}{3} & \text{if } \theta = 0.6 \\ 0 & \text{otherwise} \end{cases}$

- Applying Bayes rule with $P(D|\theta) = \binom{10}{7} \theta^7 (1-\theta)^3$ gives

\[
P(\theta|D) = \begin{cases} 0.52 & \text{if } \theta = 0.5 \\ 0.48 & \text{if } \theta = 0.6 \\ 0 & \text{otherwise} \end{cases}
\]

\[
p(\theta) = \sum \mathbb{P}(D|\theta) p(\theta)
\]

MAP estimate $\hat{\theta} = 0.5$
Bayesian inference: continuous prior example

• Suppose we have a coin of unknown probability $\theta$ of heads
  
  • We see heads 7 times in 10 tosses as the data $D$
  
  • Say we also have prior information about $\theta$: $P(\theta) = \begin{cases} 5 & \text{if } \theta \in [0.4, 0.6] \\ 0 & \text{if } \theta \not\in [0.4, 0.6] \end{cases}$

\[ P(\theta) \propto P(D|\theta)P(\theta) \]

MAP estimate $\hat{\theta} = 0.6$
Drawbacks of Bayesian inference

• Maximizing some posteriors $P(\theta|D)$ is intractable

• Some choices of prior $P(\theta)$ can overwhelm any data you observe

• It is hard to justify a choice of prior $P(\theta)$
Conjugate priors

- For a given likelihood function $P(D|\theta)$, a conjugate prior $P(\theta)$ has the following properties:
  - The prior $P(\theta)$ belongs to a family of distributions that are expressive.
  - The posterior $P(\theta|D) \propto P(D|\theta)P(\theta)$ belongs to the same family as $P(\theta)$.
  - The posterior $P(\theta|D)$ is easy to maximize.

- We will illustrate these properties for the binomial likelihood, which has a prior called the Beta distribution.
Conjugate prior is expressive

- The conjugate prior for a binomial likelihood is a $\text{Beta}(\alpha, \beta)$ distribution
  \[ P(\theta) = K(\alpha, \beta) \theta^{\alpha-1} (1 - \theta)^{\beta-1} \quad \text{where } K(\alpha, \beta) \text{ is a constant} \]

- $\text{Beta}(\alpha, \beta)$ can express a variety of shapes

- $\text{Beta}(\alpha = 1, \beta = 1)$ is uniform

Posterior is in same family as conjugate prior

- The likelihood is Binomial($N, k$) and the prior is Beta($\alpha, \beta$)

\[ P(D|\theta) = \binom{N}{k} \theta^k (1 - \theta)^{N-k} \]

\[ P(\theta) = K(\alpha, \beta) \theta^{\alpha-1} (1 - \theta)^{\beta-1} \]

- Then the posterior is Beta($\alpha + k, \beta + N - k$)

\[ P(\theta|D) = K(\alpha + k, \beta + N - k) \theta^{\alpha+k-1} (1 - \theta)^{\beta+N-k-1} \]
Updating the posterior

• Since the posterior is in the same family as the conjugate prior, the posterior can be used as a new prior if more data is observed

• Suppose we start with a uniform prior on the probability $\theta$ of heads
  • Then we observe 3H 0T
  • Then we observe 4H 3T for 7H 3T in total
  • Then we observe 10H 10T for 17H 13T in total
  • Then we observe 55H 15T for 72H 28T in total
Posterior is easy to maximize

- The posterior is $\text{Beta}(\alpha + k, \beta + N - k)$
  \[ P(\theta|D) = K(\alpha + k, \beta + N - k)\theta^{\alpha+k-1}(1 - \theta)^{\beta+N-k-1} \]

- Differentiating and setting to 0 gives the MAP estimate
  \[ \hat{\theta} = \frac{\alpha - 1 + k}{\alpha + \beta - 2 + N} \]
Conjugate priors for other likelihood functions

- If the likelihood is Bernoulli or geometric, the conjugate prior is Beta

- If the likelihood is Poisson or exponential, the conjugate prior is Gamma

- If the likelihood is normal with known variance, the conjugate prior is normal
Eigenvalues and eigenvectors review

• If $A$ is an $n \times n$ square matrix, an eigenvalue $\lambda$ and its corresponding eigenvector $\mathbf{v}$ (of dimension $n \times 1$) have the property that $A\mathbf{v} = \lambda \mathbf{v}$.

• To solve for $\lambda$, we solve the characteristic equation $|A - \lambda I| = 0$.

• Given a value of $\lambda$, we find the corresponding eigenvector(s) by solving $(A - \lambda I)\mathbf{v} = 0$.

• Note that if $\mathbf{v}$ is an eigenvector for $\lambda$, then so is any multiple $k\mathbf{v}$.
Eigenvalues and eigenvectors: example

Find the eigenvalues and eigenvectors of \( A = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix} \)

\[
|A - \lambda I| = \begin{vmatrix} 5 - \lambda & 3 \\ 3 & 5 - \lambda \end{vmatrix} = (5 - \lambda)^2 - 3^2 = \lambda^2 - 10\lambda + 25 - 9
\]

\[
= \lambda^2 - 10\lambda + 16 = (\lambda - 8)(\lambda - 2) = 0
\]

So eigenvalues are \( \lambda_1 = 8 \) and \( \lambda_2 = 2 \)
For $\lambda_1 = 8$

\[A - 8I = \begin{bmatrix} 5-8 & 3 \\ 3 & 5-8 \end{bmatrix} = \begin{bmatrix} -3 & 3 \\ 3 & -3 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}\]

So $v_1 = \begin{bmatrix} 1 \end{bmatrix}$

For $\lambda_2 = 2$

\[A - 2I = \begin{bmatrix} 5-2 & 3 \\ 3 & 5-2 \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 3 & 3 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}\]

So $v_2 = \begin{bmatrix} -1 \end{bmatrix}$
Diagonalization of a symmetric matrix

- If $A$ is an $n \times n$ symmetric square matrix, the eigenvalues are real.

- If the eigenvalues are also distinct, their eigenvectors are orthogonal.

- We can then scale the eigenvectors $v_i$ to ones of unit length $u_i$ and place them into an orthogonal matrix $U = [u_1 \ u_2 \ \ldots \ u_n]$ \( u'u = u^T \).

- Then we can write a diagonal matrix $\Lambda = U^TAU$ such that the diagonal entries of $\Lambda$ are $\lambda_1, \lambda_2, \ldots, \lambda_n$ in that order.
Diagonalization example

Diagonalize $A = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$

$\lambda_1 = 8 \Rightarrow v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow u_1 = \frac{1}{11} v_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$

$\lambda_2 = 2 \Rightarrow v_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \Rightarrow u_2 = \frac{1}{11} v_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$

$A = \begin{bmatrix} 9 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$