LECTURE 14: Introduction to Bayesian inference

- The big picture
  - motivation, applications
  - problem types (hypothesis testing, estimation, etc.)

- The general framework
  - Bayes’ rule $\rightarrow$ posterior
    (4 versions)
  - point estimates (MAP, LMS)
  - performance measures
    (prob. of error; mean squared error)
  - examples
Inference: the big picture

Real world

Predictions

Decisions

Probability theory (Analysis)

Data

Models

Inference/Statistics
Inference then and now

• Then:
  10 patients were treated: 3 died
  10 patients were not treated: 5 died
  Therefore ...

Now:

• Big data

• Big models

• Big computers
A sample of application domains

- Design and interpretation of experiments
  - polling
A sample of application domains

- marketing, advertising

- recommendation systems
  - Netflix competition
A sample of application domains

- Finance

© Source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use.
A sample of application domains

- Life sciences

  - genomics

neuroscience, etc., etc.

This image is in the public domain.
Source: Wikimedia.
A sample of application domains

- Modeling and monitoring the oceans
- Modeling and monitoring global climate
- Modeling and monitoring pollution
- Interpreting data from physics experiments
- Interpreting astronomy data
A sample of application domains

- Signal processing
  - communication systems (noisy …)
  - speech processing and understanding
  - image processing and understanding
  - tracking of objects
  - positioning systems (e.g., GPS)
  - detection of abnormal events
Model building versus inferring unobserved variables

\[ X = aS + W \]

- Model building:
  - know “signal” \( S \), observe \( X \)
  - infer \( a \)

- Variable estimation:
  - know \( a \), observe \( X \)
  - infer \( S \)
Hypothesis testing versus estimation

- Hypothesis testing:
  - unknown takes one of few possible values
  - aim at small probability of incorrect decision

  Is it an airplane or a bird?

- Estimation:
  - numerical unknown(s)
  - aim at an estimate that is "close" to the true but unknown value
The Bayesian inference framework

- Unknown $\Theta$
  - treated as a random variable
  - prior distribution $p_\Theta$ or $f_\Theta$

- Observation $X$
  - observation model $p_{X|\Theta}$ or $f_{X|\Theta}$

- Use appropriate version of the Bayes rule to find $p_{\Theta|X}(\cdot | X = x)$ or $f_{\Theta|X}(\cdot | X = x)$

- Where does the prior come from?
  - symmetry
  - known range
  - earlier studies
  - subjective or arbitrary
The output of Bayesian inference

The complete answer is a posterior distribution: PMF \( p_{\Theta|X}(\cdot | x) \) or PDF \( f_{\Theta|X}(\cdot | x) \)

Prior \( p_{\Theta} \)  

Observation Process \( x \)  

Posterior Calculation \( p_{\Theta|X}(\cdot | X = x) \)  

Point Estimates Error Analysis etc.
Point estimates in Bayesian inference

The complete answer is a posterior distribution: PMF $p_{\theta | X}(\cdot | x)$ or PDF $f_{\theta | X}(\cdot | x)$

- **Maximum a posteriori probability (MAP):**
  \[ p_{\theta | X}(\theta^* | x) = \max_{\theta} p_{\theta | X}(\theta | x), \]
  \[ f_{\theta | X}(\theta^* | x) = \max_{\theta} f_{\theta | X}(\theta | x), \]

- **Conditional expectation:** $E[\Theta | X = x]$ (LMS: Least Mean Squares)

**estimate:** $\hat{\theta} = g(x)$ (number)

**estimator:** $\hat{\Theta} = g(X)$ (random variable)
Discrete $\Theta$, discrete $X$

- values of $\Theta$: alternative hypotheses

MAP rule: $\hat{\theta} = 2$

- conditional prob of error:
  
  \[ P(\hat{\theta} \neq \Theta \mid X = x) = 0.4 \]
  
  smallest under the MAP rule

- overall probability of error:
  
  \[
  P(\hat{\Theta} \neq \Theta) = \sum_x P(\hat{\Theta} \neq \Theta \mid X = x) p_X(x) \\
  = \sum_{\theta} P(\hat{\Theta} \neq \Theta \mid \Theta = \theta) p_\Theta(\theta)
  \]
Discrete $\Theta$, continuous $X$

- Standard example:
  - send signal $\Theta \in \{1, 2, 3\}$
  
  $X = \Theta + W$

  $W \sim N(0, \sigma^2)$, indep. of $\Theta$

  $f_{X|\Theta}(x | \theta) = f_W(x - \theta)$

- conditional prob of error:

  \[
  P(\tilde{\Theta} \neq \Theta | X = x) = \sum_{\theta'} p_\Theta(\theta') f_{X|\Theta}(x | \theta')
  \]

- smallest under the MAP rule

- overall probability of error:

  \[
  P(\tilde{\Theta} \neq \Theta) = \int P(\tilde{\Theta} \neq \Theta | X = x) f_X(x) \, dx
  \]

  \[
  = \sum_{\theta} P(\tilde{\Theta} \neq \theta | \Theta = \theta) p_\Theta(\theta)
  \]

- MAP rule: $\tilde{\Theta} = 2$
Continuous $\Theta$, continuous $X$

- linear normal models
  estimation of a noisy signal

\[ X = \Theta + W \]

$\Theta$ and $W$: independent normals

multi-dimensional versions (many normal parameters, many observations)

- estimating the parameter of a uniform

\[ X: \text{uniform}[0, \Theta] \]

$\Theta$: uniform $[0, 1]$

- $\widehat{\Theta} = g(X)$
  - MAP
  - LMS

interested in:

\[ \begin{align*}
\mathbb{E}[(\widehat{\Theta} - \Theta)^2 | X = x] \\
\mathbb{E}[(\Theta - \Theta)^2]
\end{align*} \]
Inferring the unknown bias of a coin and the Beta distribution

- Standard example:
  - coin with bias $\Theta$; prior $f_\Theta(\cdot)$
  - fix $n$; $K =$ number of heads

- Assume $f_\Theta(\cdot)$ is uniform in $[0, 1]$

  $f_{\Theta|K}(\theta | k) = \frac{1 \cdot \binom{m}{k} \theta^k (1 - \theta)^{m-k}}{p_K(k)}$

  $= \frac{1}{d(n, k)} \theta^k (1 - \theta)^{n-k}$

  “Beta distribution, with parameters $(k + 1, n - k + 1)$”

- If prior is Beta: $f_\Theta(\theta) = \frac{1}{c} \theta^\alpha (1 - \theta)^\beta$
  $\alpha, \beta > 0$

  $f_{\Theta|K}(\theta | k) = \frac{1}{c} \theta^\alpha (1 - \theta)^\beta \binom{m}{k} \theta^k (1 - \theta)^{m-k} p_K(k) = d \theta^{\alpha+k} (1 - \theta)^{\beta+n-k}$
Inferring the unknown bias of a coin: point estimates

- Standard example:
  - coin with bias \( \Theta \); prior \( f_\Theta(\cdot) \)
  - fix \( n \); \( K \) = number of heads
- Assume \( f_\Theta(\cdot) \) is uniform in \([0, 1]\)

\[
f_\Theta|K(\theta | k) = \frac{1}{d(n,k)} \theta^k (1-\theta)^{n-k}
\]

- MAP estimate:

\[
\hat{\theta}_{\text{MAP}} = \frac{k}{n}
\]

\[
\Theta_{\text{MAP}} = \frac{K}{n}
\]

\[
\int_0^1 \theta^\alpha (1-\theta)^\beta d\theta = \frac{\alpha! \beta!}{(\alpha + \beta + 1)!}
\]

\[
E[\Theta | K = k] = \int_0^1 \theta f_\Theta|K(\theta | k) d\theta
\]

\[
= \frac{1}{d(n,k)} \int_0^1 \theta^{k+1} (1-\theta)^{n-k} d\theta
\]

\[
= \frac{(k+1)! (n-k)!}{(n+2)!} \cdot \frac{k+1}{n+2} \cdot \frac{1}{K_0 (n-k)^!}
\]

\[
\approx \frac{k}{n} \quad \text{when } n \text{ large}
\]
Summary

- Problem data: $p_\Theta(\cdot), p_{X|\Theta}(\cdot | \cdot)$

- Given the value $x$ of $X$: find, e.g., $p_{\Theta|X}(\cdot | x)$
  - using appropriate version of the Bayes rule (4 choices)

- Estimator $\hat{\Theta} = g(X)$
  - MAP: $\hat{\theta}_{\text{MAP}} = g_{\text{MAP}}(x)$ maximizes $p_{\Theta|X}(\theta | x)$
  - LMS: $\hat{\theta}_{\text{LMS}} = g_{\text{LMS}}(x) = \mathbb{E}[\Theta | X = x]$