# Bayesian Methods

David Rosenberg

New York University

April 20, 2016

# Classical Statistics

# Frequentist or "Classical" Statistics

- Probability model with parameter $\theta \in \Theta$

$$\{p(y;\theta) \mid \theta \in \Theta\},$$

  where $p(y;\theta)$ is either a PDF or a PMF.
- Assume that $p(y;\theta)$ governs the world we are observing.
- In **frequentist statistics**, the **parameter** $\theta$ is a
  - **fixed constant** (i.e. not random) and is
  - **unknown** to us.
- If we knew $\theta$, there would be no need for statistics.
- Instead of $\theta$, we have a **sample** $\mathcal{D} = \{y_1, \ldots, y_n\}$ i.i.d. $p(y;\theta)$.
- Statistics is about how to use $\mathcal{D}$ in place of $\theta$.

# Point Estimation

- One type of statistical problem is **point estimation**.
- A **statistic** $s = s(\mathcal{D})$ is any function of the data.
- A statistic $\hat{\theta} = \hat{\theta}(\mathcal{D})$ is a **point estimator** if $\hat{\theta} \approx \theta$.
- Desirable statistical properties of point estimators:
  - **Consistency:** As data size $n \to \infty$, we get $\hat{\theta} \to \theta$.
  - **Efficiency:** (Roughly speaking) $\hat{\theta}_n$ is as accurate as we can get from a sample of size $n$.
  - e.g. **maximum likelihood estimation** is consistent and efficient under reasonable conditions.
- In frequentist statistics, you can make up any estimator you want.
  - Justify its use by showing it has desirable properties.

# Bayesian Statistics: Introduction

# Bayesian Statistics

- Major viewpoint change in **Bayesian statistics**:
  - parameter $\theta \in \Theta$ is a **random variable**.
- New ingredient is the **prior distribution:**
  - It is a distribution on parameter space $\Theta$.
  - Reflects our belief about $\theta$.
  - Must be chosen before seeing any data.

# The Bayesian Method

1. Define the model:
   - Choose a distribution $p(\theta)$, called the **prior distribution**.
   - Choose a probability model or "**likelihood model**", now written as:

$$\{p(\mathcal{D} \mid \theta) \mid \theta \in \Theta\}.$$

2. After observing $\mathcal{D}$, compute the **posterior distribution** $p(\theta \mid \mathcal{D})$.
3. Choose **action** based on $p(\theta \mid \mathcal{D})$.

# The Posterior Distribution

- By Bayes rule, can write the posterior distribution as

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})}.$$

- **likelihood:** $p(\mathcal{D} \mid \theta)$
- **prior:** $p(\theta)$
- **marginal likelihood**: $p(\mathcal{D})$.
- Note: $p(\mathcal{D})$ is just a normalizing constant for $p(\theta \mid \mathcal{D})$. Can write

$$\underbrace{p(\theta \mid \mathcal{D})}_{\text{posterior}} \propto \underbrace{p(\mathcal{D} \mid \theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}.$$

# Recap and Interpretation

- Prior represents belief about $\theta$ before observing data $\mathcal{D}$.
- Posterior represents the **rationally "updated" beliefs** after seeing $\mathcal{D}$.
- All inferences and action-taking are based on the posterior distribution.
- In the Bayesian approach,
  - No issue of "choosing a procedure" or justifying an estimator.
  - Only choices are the **prior** and the **likelihood model**.
  - For decision making, need a **loss function**.
  - Everything after that is **computation**.

# Coin Flipping: The Beta-Binomial Model

# Coin Flipping: Setup

- **Parameter space** $\theta \in \Theta = [0,1]$:

$$\mathbb{P}(\text{Heads} \mid \theta) = \theta.$$

- Data $\mathcal{D} = \{H, H, T, T, T, T, T, H, \ldots, T\}$
  - $n_h$: number of heads
  - $n_t$: number of tails
- **Likelihood model** (Bernoulli Distribution):

$$p(\mathcal{D} \mid \theta) = \theta^{n_h} (1-\theta)^{n_t}$$

  - (probability of getting the flips in the order they were received)

# Coin Flipping: Beta Prior

- **Prior:**

$$
\begin{aligned}
\theta &\sim \text{Beta}(\alpha, \beta) \\
p(\theta) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}
\end{aligned}
$$


Figure by Horas based on the work of Krishnavedala (Own work) [Public domain], via Wikimedia Commons http://commons.wikimedia.org/wiki/File:Beta_distribution_pdf.svg.

# Coin Flipping: Beta Prior

- **Prior:**

$$\begin{aligned}\theta &\sim \text{Beta}(h,t)\\ p(\theta) &\propto \theta^{h-1}(1-\theta)^{t-1}\end{aligned}$$

- **Mean of Beta distribution:**

$$\mathbb{E}\theta = \frac{h}{h+t}$$

# Coin Flipping: Posterior

- **Prior:**

$$\begin{aligned}\theta &\sim \text{Beta}(h,t)\\ p(\theta) &\propto \theta^{h-1}(1-\theta)^{t-1}\end{aligned}$$

- **Likelihood model**:

$$p(\mathcal{D}\mid\theta) = \theta^{n_h}(1-\theta)^{n_t}$$

- **Posterior density:**

$$\begin{aligned}p(\theta\mid\mathcal{D}) &\propto p(\theta)p(\mathcal{D}\mid\theta)\\ &\propto \theta^{h-1}(1-\theta)^{t-1}\times\theta^{n_h}(1-\theta)^{n_t}\\ &= \theta^{h-1+n_h}(1-\theta)^{t-1+n_t}\end{aligned}$$

# Posterior is Beta

- **Prior:**

$$\begin{aligned} \theta &\sim \text{Beta}(h,t) \\ p(\theta) &\propto \theta^{h-1}(1-\theta)^{t-1} \end{aligned}$$

- **Posterior density:**

$$p(\theta \mid \mathcal{D}) \propto \theta^{h-1+n_h}(1-\theta)^{t-1+n_t}$$

- **Posterior is in the beta family**:

$$\theta \mid \mathcal{D} \sim \text{Beta}(h+n_h, t+n_t)$$

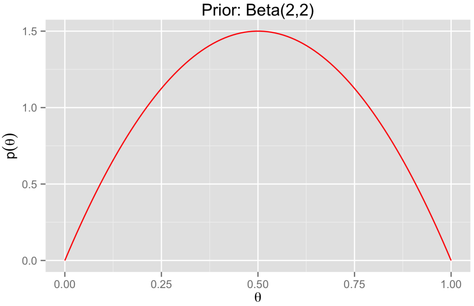
- **Interpretation**:
  - Prior initializes our counts with $h$ heads and $t$ tails.
  - Posterior increments counts by observed $n_h$ and $n_t$.

# Example: Coin Flipping

- Suppose we have a coin, possibly biased

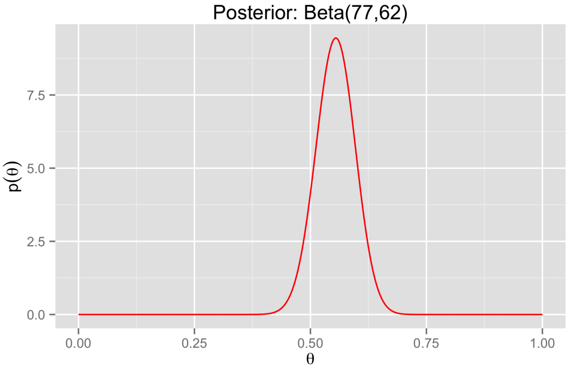
$$\mathbb{P}(\text{Heads} \mid \theta) = \theta.$$

- **Parameter space** $\theta \in \Theta = [0,1]$.
- **Prior distribution:** $\theta \sim \text{Beta}(2,2)$.



# Example: Coin Flipping

- Next, we gather some data $\mathcal{D} = \{H, H, T, T, T, T, T, H, \ldots, T\}$:
- Heads: 75  Tails: 60
  - $\hat{\theta}_{\text{MLE}} = \frac{75}{75+60} \approx 0.556$
- **Posterior distribution:** $\theta \mid \mathcal{D} \sim \textbf{Beta}(77, 62)$:



# Bayesian Point Estimates

- Suppose we have posterior $\theta \mid \mathcal{D}$...
- But we want a point estimate $\hat{\theta}$ or $\theta$.
- Common options:
  - **posterior mean** $\hat{\theta} = \mathbb{E}[\theta \mid \mathcal{D}]$
  - **maximum a posteriori (MAP) estimate** $\hat{\theta} = \arg\max_{\theta} p(\theta \mid \mathcal{D})$
    - Note: this is the **mode** of the posterior distribution

# What else can we do with a posterior?

- Look at it.
- Extract "**credible set**" for $\theta$ (a Bayesian confidence interval).
    - e.g. Interval $[a, b]$ is a 95% **credible set** if

$$\mathbb{P}\left(\theta \in [a, b] \mid \mathcal{D}\right) \geqslant 0.95$$

- The most "Bayesian" approach is **Bayesian decision theory**:
    - Choose a loss function.
    - Find action **minimizing expected risk w.r.t. posterior**

# Bayesian Regression

# Bayesian Conditional Models

- Input space $\mathcal{X} = \mathbf{R}^d$ Output space $\mathcal{Y} = \mathbf{R}$
- **Conditional probability model**, or **likelihood model**:

$$\{p(y \mid x, \theta) \mid \theta \in \Theta\}$$

  - Conditional here refers to the conditioning on the input $x$.
  - Means that $x$'s are known and not governed by our probability model.

# Gaussian Regression Model

- Input space $\mathcal{X} = \mathbf{R}^d$ Output space $\mathcal{Y} = \mathbf{R}$
- **Conditional probability model**, or **likelihood model**:

$$y \mid x, \theta \sim \mathcal{N}\left(\theta^T x, \sigma^2\right),$$

  for some known $\sigma^2 > 0$.
- **Parameter space** $\Theta = \mathbf{R}^d$.
- **Data:** $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
  - Write $y = (y_1, \ldots, y_n)$ and $x = (x_1, \ldots, x_n)$.
  - Assume $y_i$'s are **conditionally independent**, given $x$ and $\theta$.

# Gaussian Likelihood

- The **likelihood** of $\theta \in \Theta$ for the data $\mathcal{D}$ is

$$
\begin{aligned}
p(y \mid x, \theta) &= \prod_{i=1}^{n} p(y_i \mid x_i, \theta) \qquad \text{by conditional independence.} \\
&= \prod_{i=1}^{n} \left[ \frac{1}{\sigma\sqrt{2\pi}} \exp\left( -\frac{(y_i - \theta^T x_i)^2}{2\sigma^2} \right) \right]
\end{aligned}
$$

- Recall from a previous lecture[1] that the **MLE** is

$$
\begin{aligned}
\theta^*_{\text{MLE}} &= \underset{\theta \in \mathbf{R}^d}{\arg\max}\, p(y \mid x, \theta) \\
&= \underset{\theta \in \mathbf{R}^d}{\arg\min} \sum_{i=1}^{n} (y_i - \theta^T x_i)^2
\end{aligned}
$$

[1] `https://davidrosenberg.github.io/ml2015/docs/8.Lab.glm.pdf`, slide 5.

# Priors and Posteriors

- Choose a Gaussian **prior distribution** $p(\theta)$ on $\Theta$:

$$\theta \sim \mathcal{N}(0, \Sigma_0)$$

  for some **covariance matrix** $\Sigma_0 \succ 0$ (i.e. $\Sigma_0$ is spd).

- **Posterior distribution**

$$\begin{aligned} p(\theta \mid \mathcal{D}) &= p(\theta \mid x, y) \\ &= p(y \mid x, \theta)\, p(\theta) / p(y) \\ &\propto p(y \mid x, \theta) p(\theta) \\ &= \prod_{i=1}^{n} \left[ \frac{1}{\sigma\sqrt{2\pi}} \exp\left( -\frac{(y_i - \theta^T x_i)^2}{2\sigma^2} \right) \right] \quad \textbf{(likelihood)} \\ &\quad \times |2\pi\Sigma_0|^{-1/2} \exp\left( -\frac{1}{2} \theta^T \Sigma_0^{-1} \theta) \right) \quad \textbf{(prior)} \end{aligned}$$

# Example in 1-Dimension

- Input space $\mathcal{X} = [-1, 1]$ Output space $\mathcal{Y} = \mathbf{R}$
- Basic Gaussian regression model:

$$y = w_0 + w_1 x + \varepsilon,$$

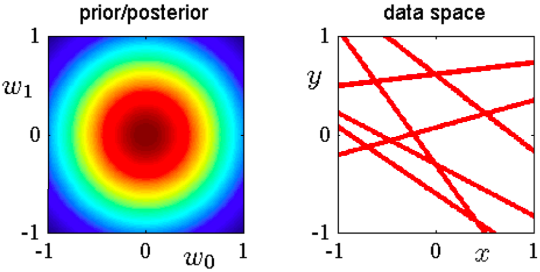
where $\varepsilon \sim \mathcal{N}(0, 0.2^2)$.

- Written another way, the **likelihood model** is

$$y \mid x, w_0, w_1 \quad \sim \quad \mathcal{N}\left(w_0 + w_1 x\,,\, 0.2^2\right).$$

# Example in 1-Dimension

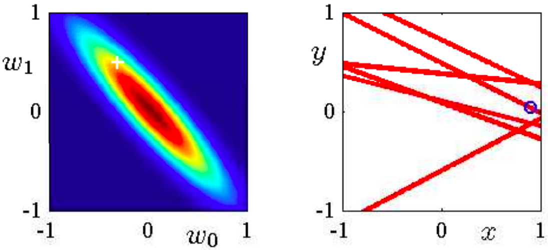
- **Prior distribution:** $\theta = (w_0, w_1) \sim \mathcal{N}\left(0, \frac{1}{2}I\right)$




- On right, plots of $y = w_0 + w_1 x$ for random $(w_0, w_1) \sim p(\theta) = \mathcal{N}(0, \frac{1}{2}I)$.

Bishop's PRML Fig 3.7

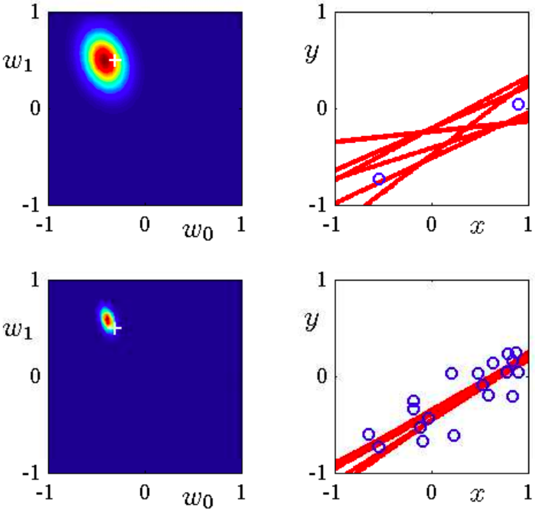
# Example in 1-Dimension: 1 Observation




- On left, posterior distribution; white '+' indicates true parameter values
- On right, the blue circle indicates the training observation

Bishop's PRML Fig 3.7

# Example in 1-Dimension: 2 and 20 Observations




Bishop's PRML Fig 3.7

# Closed Form for Posterior

- Model:

$$\begin{aligned} \theta &\sim \mathcal{N}(0, \Sigma_0) \\ y_i \mid x, \theta &\text{ i.i.d. } \mathcal{N}(\theta^T x_i, \sigma^2) \end{aligned}$$

- Design matrix $X$ Response column vector $y$
- **Posterior distribution is a Gaussian distribution:**

$$\begin{aligned} \theta \mid \mathcal{D} &\sim \mathcal{N}(\mu_P, \Sigma_P) \\ \boldsymbol{\Sigma_P} &= \left(\sigma^{-2} X^T X + \Sigma_0^{-1}\right)^{-1} \\ \boldsymbol{\mu_P} &= \left(X^T X + \sigma^2 \Sigma_0^{-1}\right)^{-1} X^T y \end{aligned}$$

- **Posterior Variance $\Sigma_P$** gives us a natural **uncertainty measure.**

See Rasmussen and Williams' *Gaussian Processes for Machine Learning*, Ch 2.1.
http://www.gaussianprocess.org/gpml/chapters/RW2.pdf

# Closed Form for Posterior

- **Posterior distribution is a Gaussian distribution:**

$$
\begin{aligned}
\theta \mid \mathcal{D} &\sim \mathcal{N}(\mu_P, \Sigma_P) \\
\Sigma_P &= \left(\sigma^{-2} X^T X + \Sigma_0^{-1}\right)^{-1} \\
\mu_P &= \sigma^{-2} \Sigma_P X^T y
\end{aligned}
$$

- The **MAP estimator** and the **posterior mean** are given by

$$\mu_P = \left(X^T X + \sigma^2 \Sigma_0^{-1}\right)^{-1} X^T y$$

- Look familiar?
- For the prior variance $\Sigma_0 = \frac{\sigma^2}{\lambda} I$, we get

$$\mu_P = \left(X^T X + \lambda I\right)^{-1} X^T y,$$

which is of course the ridge regression solution.

# Posterior Mean and Posterior Mode (MAP)

- **Posterior density for $\Sigma_0 = \frac{\sigma^2}{\lambda} I$:**

$$p(\theta \mid \mathcal{D}) \quad \propto \quad \underbrace{\exp\left(-\frac{\lambda}{2\sigma^2}\|\theta\|^2\right)}_{\textbf{prior}} \underbrace{\prod_{i=1}^{n} \exp\left(-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}\right)}_{\textbf{likelihood}}$$

- **To find MAP, sufficient to minimize the negative log posterior:**

$$\begin{aligned} \hat{\theta}_{\mathbf{MAP}} &= \underset{\theta \in \mathbf{R}^d}{\arg\min}\left[-\log p(\theta \mid \mathcal{D})\right] \\ &= \underset{\theta \in \mathbf{R}^d}{\arg\min} \underbrace{\sum_{i=1}^{n}(y_i - \theta^T x_i)^2}_{\textbf{log-likelihood}} + \underbrace{\lambda\|\theta\|^2}_{\textbf{log-prior}} \end{aligned}$$

- Which is the ridge regression objective.

# Predictive Distribution

- Given a new input point $x_{\text{new}}$, how to predict $y_{\text{new}}$ ?
- **Predictive distribution**

$$
\begin{aligned}
& p(y_{\mathbf{new}} \mid x_{\mathbf{new}}, \mathcal{D}) \\
= & \int p(y_{\mathbf{new}} \mid x_{\mathbf{new}}, \theta, \mathcal{D}) p(\theta \mid \mathcal{D})\, d\theta \\
= & \int p(y_{\mathbf{new}} \mid x_{\mathbf{new}}, \theta) p(\theta \mid \mathcal{D})\, d\theta
\end{aligned}
$$

- For Gaussian regression, posterior and predictive distributions have closed forms.

# Closed Form for Predictive Distribution

- **Model**:

$$\begin{aligned} \theta &\sim \mathcal{N}(0, \Sigma_0) \\ y_i \mid x, \theta &\text{ i.i.d. } \mathcal{N}(\theta^T x_i, \sigma^2) \end{aligned}$$

- **Predictive Distribution**

$$p(y_{\text{new}} \mid x_{\text{new}}, \mathcal{D}) = \int p(y_{\text{new}} \mid x_{\text{new}}, \theta) p(\theta \mid \mathcal{D}) \, d\theta.$$

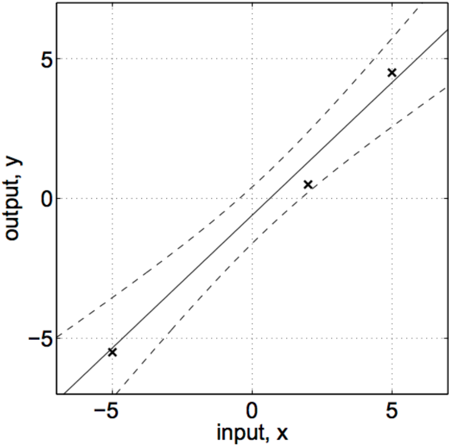
  - Averages over prediction for each $\theta$, weighted by posterior distribution.

- **Closed form:**

$$\begin{aligned} y_{\text{new}} \mid x_{\text{new}}, \mathcal{D} &\sim \mathcal{N}(\eta_{\text{new}}, \sigma_{\text{new}}) \\ \mu_{\text{new}} &= \mu_P^T x_{\text{new}} \\ \sigma_{\text{new}} &= \underbrace{x_{\text{new}}^T \Sigma_P x_{\text{new}}}_{\text{from variance in } \theta} + \underbrace{\sigma^2}_{\text{inherent variance in } y} \end{aligned}$$

# Predictive Distributions

- With predictive distributions, can draw error bands:




Rasmussen and Williams' *Gaussian Processes for Machine Learning*, Fig.2.1(b)