

# Bayesian Methods

David Rosenberg

New York University

October 29, 2016

# 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**  $\mathcal{D} = \{y_1, \dots, y_n\}$  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  $n \rightarrow \infty$ , we get  $\hat{\theta} \rightarrow \theta$ .
  - **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

- 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

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

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

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

# Recap and Interpretation

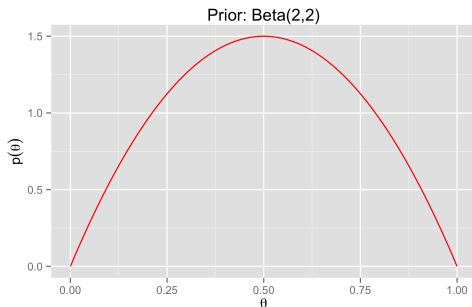
- Prior represents belief about  $\theta$  before observing data  $\mathcal{D}$ .
- Posterior represents the **rationally “updated” beliefs** after seeing  $\mathcal{D}$ .
- All inferences and action-taking are based on the posterior distribution.
- In the Bayesian approach,
  - No issue of “choosing 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**.

## Example: Coin Flipping

- Suppose we have a coin, possibly biased

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

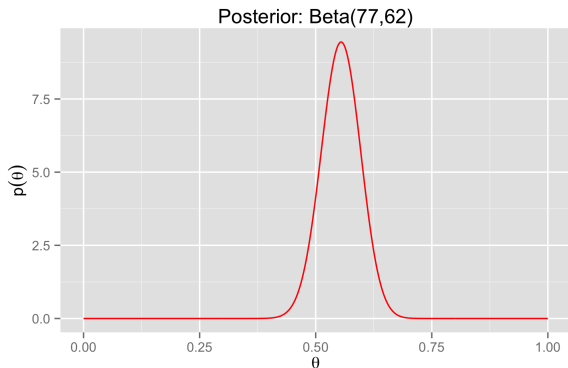
- **Parameter space**  $\theta \in \Theta = [0, 1]$ .
- **Prior distribution:**  $\theta \sim \text{Beta}(2, 2)$ .





## Example: Coin Flipping

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



# What to do with the Posterior Distribution?

- Look at it.
- Extract a point estimate of  $\theta$  (e.g. mean or mode of posterior).
- Extract “**credible set**” for  $\theta$  (a Bayesian confidence interval).
  - e.g. Interval  $[a, b]$  is a 95% **credible set** if

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

- The most “Bayesian” approach is **Bayesian decision theory**:
  - Choose a loss function.
  - Find action minimizing “posterior risk”.

# Bayesian Decision Theory

- Ingredients:
  - **Action space**  $\mathcal{A}$ .
  - **Parameter space**  $\Theta$ .
  - **Loss function**:  $\ell : \mathcal{A} \times \Theta \rightarrow \mathbf{R}$ .
  - **Prior**: Distribution  $p(\theta)$  on  $\Theta$ .
- The **posterior risk** of an action  $a \in \mathcal{A}$  is

$$\begin{aligned} r(a) &:= \mathbb{E}[\ell(\theta, a) \mid \mathcal{D}] \\ &= \int \ell(\theta, a) p(\theta \mid \mathcal{D}) d\theta. \end{aligned}$$

- It's the **expected loss under the posterior**.
- A **Bayes action**  $a^*$  is an action that minimizes posterior risk:

$$r(a^*) = \min_{a \in \mathcal{A}} r(a)$$

# Bayesian Point Estimation

- General Setup:
  - Data  $\mathcal{D}$  generated by  $p(y | \theta)$ , for unknown  $\theta \in \Theta$ .
  - Want to produce a **point estimate** for  $\theta$ .
- Choose the following:
  - **Loss**  $\ell(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2$
  - **Prior**  $p(\theta)$  on  $\Theta$ .
- Find **action**  $\hat{\theta} \in \Theta$  that minimizes **posterior risk**:

$$\begin{aligned}
 r(\hat{\theta}) &= \mathbb{E} \left[ (\theta - \hat{\theta})^2 \mid \mathcal{D} \right] \\
 &= \int (\theta - \hat{\theta})^2 p(\theta \mid \mathcal{D}) d\theta
 \end{aligned}$$

# Bayesian Point Estimation: Square Loss

- Find **action**  $\hat{\theta} \in \Theta$  that minimizes **posterior risk**

$$r(\hat{\theta}) = \int (\theta - \hat{\theta})^2 p(\theta | \mathcal{D}) d\theta.$$

- Differentiate:

$$\begin{aligned} \frac{dr(\hat{\theta})}{d\hat{\theta}} &= -\int 2(\theta - \hat{\theta}) p(\theta | \mathcal{D}) d\theta \\ &= -2 \int \theta p(\theta | \mathcal{D}) d\theta + 2\hat{\theta} \underbrace{\int p(\theta | \mathcal{D}) d\theta}_{=1} \\ &= -2 \int \theta p(\theta | \mathcal{D}) d\theta + 2\hat{\theta} \end{aligned}$$

# Bayesian Point Estimation: Square Loss

- Derivative of posterior risk is

$$\frac{dr(\hat{\theta})}{d\hat{\theta}} = -2 \int \theta p(\theta | \mathcal{D}) d\theta + 2\hat{\theta}.$$

- First order condition  $\frac{dr(\hat{\theta})}{d\hat{\theta}} = 0$  gives

$$\begin{aligned}\hat{\theta} &= \int \theta p(\theta | \mathcal{D}) d\theta \\ &= \mathbb{E}[\theta | \mathcal{D}]\end{aligned}$$

- Bayes action for square loss is the posterior mean.**

# Bayesian Point Estimation: Absolute Loss

- **Loss:**  $\ell(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$
- **Bayes action for absolute loss is the posterior median.**
  - That is, the median of the distribution  $p(\theta | \mathcal{D})$ .
  - Show with approach similar to what was used in Homework #1.

# Bayesian Point Estimation: Zero-One Loss

- Suppose  $\Theta$  is discrete (e.g.  $\Theta = \{\text{english}, \text{french}\}$ )
- **Zero-one loss:**  $\ell(\theta, \hat{\theta}) = 1(\theta \neq \hat{\theta})$
- **Posterior risk:**

$$\begin{aligned}
 r(\hat{\theta}) &= \mathbb{E} \left[ 1(\theta \neq \hat{\theta}) \mid \mathcal{D} \right] \\
 &= \mathbb{P}(\theta \neq \hat{\theta} \mid \mathcal{D}) \\
 &= 1 - \mathbb{P}(\theta = \hat{\theta} \mid \mathcal{D}) \\
 &= 1 - p(\hat{\theta} \mid \mathcal{D})
 \end{aligned}$$

- **Bayes action is**

$$\hat{\theta} = \arg \max_{\theta \in \Theta} p(\theta \mid \mathcal{D})$$

- This  $\hat{\theta}$  is called the **maximum a posteriori (MAP)** estimate.
- The MAP estimate is the **mode** of the posterior distribution.



# Bayesian Point Estimation: Custom Loss Function

- Suppose  $\Theta$  is discrete (e.g.  $\Theta = \{\text{english}, \text{french}\}$ )
- **Loss function**  $\ell(\hat{\theta}, \theta)$ :

$$\ell(\text{french}, \text{english}) = 10$$

$$\ell(\text{english}, \text{french}) = 1$$

$$\ell(\text{english}, \text{english}) = 0$$

$$\ell(\text{french}, \text{french}) = 0$$

- **Posterior risk:**

$$r(\text{french}) = 10p(\text{english} \mid \mathcal{D}) + 0p(\text{french} \mid \mathcal{D})$$

$$r(\text{english}) = 1p(\text{french} \mid \mathcal{D}) + 0p(\text{english} \mid \mathcal{D})$$

- **Bayes action** is french iff  $r(\text{french}) < r(\text{english})$ , i.e.

$$p(\text{english} \mid \mathcal{D}) < \frac{1}{10}p(\text{french} \mid \mathcal{D}).$$

# 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:**

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

for some known  $\sigma^2 > 0$ .

- **Parameter space**  $\Theta = \mathbf{R}^d$ .
- **Data:**  $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ 
  - Write  $y = (y_1, \dots, y_n)$  and  $x = (x_1, \dots, x_n)$ .
  - 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

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

- Recall from the GLM lecture<sup>1</sup> that the **MLE** is

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

<sup>1</sup><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**

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

## Example in 1-Dimension

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

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

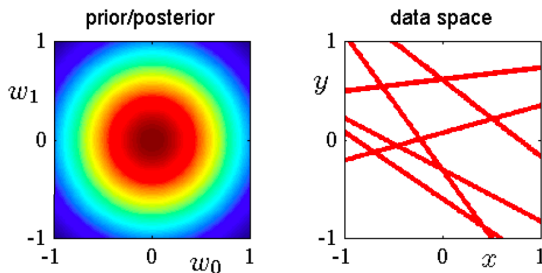
where  $\varepsilon \sim \mathcal{N}(0, 0.2^2)$ .

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

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

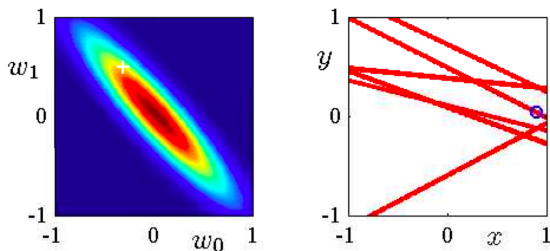
# Example in 1-Dimension

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



- On right, plots of  $y = w_0 + w_1x$  for random  $(w_0, w_1) \sim p(\theta) = \mathcal{N}(0, \frac{1}{2}I)$ .

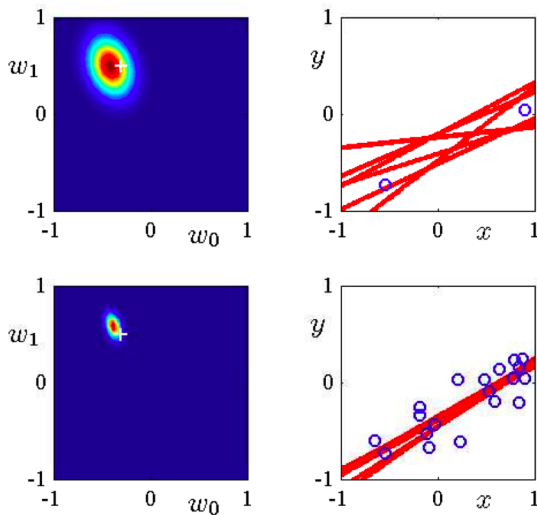
# Example in 1-Dimension: 1 Observation



- On left, the white cross indicates the true parameter values.
- On right, the blue circle indicates the training observation.



# Example in 1-Dimension: 2 and 20 Observations



Bishop's PRML Fig 3.7

# Predictive Distribution

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

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

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

# Closed Form for Posterior

- Model:

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

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

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

- 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:**

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

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

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

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

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

which is of course the ridge regression solution.

# Posterior Mean and Posterior Mode (MAP)

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

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

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

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

- Which is the ridge regression objective.

# Closed Form for Predictive Distribution

- Model:

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

- Predictive Distribution

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

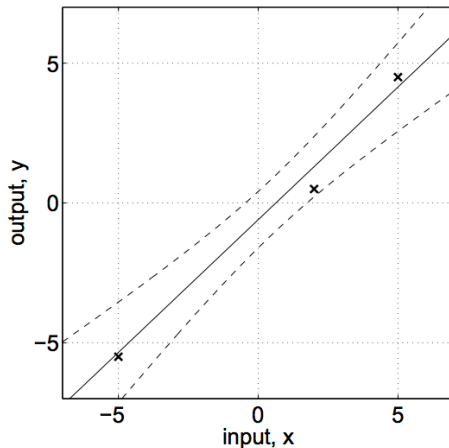
- Averages over prediction for each  $\theta$ , weighted by posterior distribution.

- Closed form:

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

# Predictive Distributions

- With predictive distributions, can draw error bands:



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

# Bayesian Predictive Distributions vs GLMs

- Gaussian regression with MLE, from our GLM lecture:
  - produces a Gaussian for each input  $x$ .

$$x \mapsto \mathcal{N}(x^T \theta_{\text{MLE}}, \sigma^2)$$

- Bayesian predictive distributions:
  - produce a Gaussian for each input  $x$

$$x \mapsto \mathcal{N} \left( \theta_{\text{ridge}}^T x, \underbrace{x_{\text{new}}^T \Sigma_P x_{\text{new}}}_{\text{from variance in } \theta} + \underbrace{\sigma^2}_{\text{inherent variance in } y} \right)$$

- In Bayesian version
  - equivalent to using a **regularized** least squares fit
  - variance has additional piece from uncertainty in  $\theta$



# Conjugate Prior Examples

- A prior is conjugate for a likelihood model if the posterior is in the same “family” as the prior.
- ① If prior is a beta distribution, and likelihood model is a Bernoulli distribution, then posterior is a beta distribution.
  - Prior and posterior in the same family  $\implies$  **Beta is a conjugate prior for Bernoulli**
- ② If prior is a Gaussian distribution, and likelihood model is a Gaussian distribution, then posterior is a Gaussian distribution.
  - Prior and posterior in the same family  $\implies$  **Gaussian is a conjugate prior for Gaussian**

Conjugacy of the prior is really a statement about the prior **family**.

# Conjugate Prior Family

- Let  $\pi$  be a family of prior distributions on  $\Theta$ .
- Let  $P$  be likelihood model with parameter space  $\Theta$ .

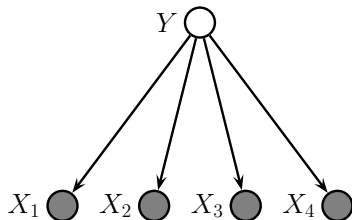
## Definition

A family of distributions  $\pi$  is **conjugate to** likelihood model  $P$  if for any prior in  $\pi$ , the posterior is always in  $\pi$ .

- Trivial Example:
  - The family of all probability distributions is conjugate to any likelihood model.
- Every exponential family has a nontrivial conjugate prior family. (KPM Section 9.2)

# Naive Bayes: A Generative Model for Classification

- $\mathcal{X} = \left\{ (X_1, X_2, X_3, X_4) \in \{0, 1\}^4 \right\}$        $\mathcal{Y} = \{0, 1\}$  be a class label.
- Consider the Bayesian network depicted below:



- BN structure implies joint distribution factors as:

$$p(x_1, x_2, x_3, x_4, y) = p(y)p(x_1 | y)p(x_2 | y)p(x_3 | y)p(x_4 | y)$$

- Features  $X_1, \dots, X_4$  are independent given the class label  $Y$ .

# Parameterized Expression for Joint Distribution

- Parameters:**

$$\mathbb{P}(Y = 1) = \theta_y \quad \mathbb{P}(X_i = 1 \mid Y = 1) = \theta_{i1} \quad \mathbb{P}(X_i = 1 \mid Y = 0) = \theta_{i0}$$

- Joint distribution is**

$$\begin{aligned} & p(x_1, \dots, x_d, y) \\ = & p(y) \prod_{i=1}^n p(x_i \mid y) \\ = & (\theta_y)^y (1 - \theta_y)^{1-y} \\ & \times \prod_{i=1}^n (\theta_{i1})^{yx_i} (1 - \theta_{i1})^{y(1-x_i)} (\theta_{i0})^{(1-y)x_i} (1 - \theta_{i0})^{(1-y)(1-x_i)} \end{aligned}$$

# Maximum Likelihood Estimators for Naive Bayes

- Training set  $\mathcal{D} = \{(x^1, y^1), \dots, (x^n, y^n)\}$ .
- Obvious “plug-in” estimators for the Naive Bayes model are also MLEs:

$$\mathbb{P}(Y = 1) \approx \hat{\theta}_y = \frac{1}{n} \sum_{i=1}^n 1(y^i = 1)$$

$$\mathbb{P}(X_i = 1 \mid Y = 1) \approx \hat{\theta}_{i1} = \frac{\sum_{j=1}^n 1(y^j = 1 \text{ and } x_i^j = 1)}{\sum_{j=1}^n 1(y^j = 1)}$$

$$\mathbb{P}(X_i = 1 \mid Y = 0) = \hat{\theta}_{i0} = \frac{\sum_{j=1}^n 1(y^j = 0 \text{ and } x_i^j = 1)}{\sum_{j=1}^n 1(y^j = 0)}$$

## Example: SPAM Classification

- Label  $Y \in \mathcal{Y} = \{\text{SPAM}, \text{HAM}\}$ .
- Features  $X_i \in \{0, 1\}$ .
- Bag of words representation:

$$X_i = 1(\text{Email contains word "Private\_Jet"})$$

- After parameter estimation, prediction done with

$$p(\text{SPAM} | x) \propto p(\text{SPAM}) \prod_{i=1}^d \hat{p}(x_i | \text{SPAM}).$$

- Each  $\hat{p}(x_i | \text{SPAM})$  is the estimated probability that  $x_i$  would be observed (or not) in a SPAM message.
- Issue: What if we never see  $X_1 = 1$  when  $Y = \text{SPAM}$  in  $\mathcal{D}$ ?
  - Then whenever we see  $X_1 = 1$ , we will predict  $p(\text{SPAM} | x) = 0$ .