Probabilistic models

-

Bayesian Methods

CDS, NYU

March 21, 2023

Contents

Table of Contents

Parametric Family of Densities

- A **parametric family of densities** is a set

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

  - where $p(y \mid \theta)$ is a density on a **sample space** $\mathcal{Y}$, and
  - θ is a **parameter** in a [finite dimensional] **parameter space** Θ.
- This is the common starting point for a treatment of classical or Bayesian statistics.
- In this lecture, whenever we say "density", we could replace it with "mass function." (and replace integrals with sums).

Frequentist or "Classical" Statistics

- We're still working with a parametric family of densities:

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

- Assume that $p(y \mid \theta)$ governs the world we are observing, for some $\theta \in \Theta$.
- If we knew the right $\theta \in \Theta$, there would be no need for statistics.
- But instead of θ, we have data $\mathcal{D}$: $y_1, \ldots, y_n$ sampled i.i.d. from $p(y \mid \theta)$.
- Statistics is about how to get by with $\mathcal{D}$ in place of θ.

Point Estimation

- One type of statistical problem is **point estimation**.
- A **statistic** $s = s(\mathcal{D})$ is any function of the data.
- A statistic $\hat{\theta} = \hat{\theta}(\mathcal{D})$ taking values in Θ is a **point estimator of** θ.
- A good point estimator will have $\hat{\theta} \approx \theta$.
- **Desirable statistical properties of point estimators**:
  - **Consistency:** As data size $n \to \infty$, we get $\hat{\theta}_n \to \theta$.
  - **Efficiency:** (Roughly speaking) $\hat{\theta}_n$ is as accurate as we can get from a sample of size n.
- **Maximum likelihood estimators** are consistent and efficient under reasonable conditions.

Example of Point Estimation: Coin Flipping

- Parametric family of mass functions:

$$p(\text{Heads} \mid \theta) = \theta,$$

for $\theta \in \Theta = (0,1)$.

Coin Flipping: MLE

- Data $\mathcal{D} = (H, H, T, T, T, T, T, H, \ldots, T)$, assumed i.i.d. flips.
  - n_h: number of heads
  - n_t: number of tails
- **Likelihood function** for data $\mathcal{D}$:

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

- As usual, it is easier to maximize the log-likelihood function:

$$\begin{aligned}\hat{\theta}_{\text{MLE}} &= \underset{\theta \in \Theta}{\arg\max} \log L_{\mathcal{D}}(\theta) \\ &= \underset{\theta \in \Theta}{\arg\max} \left[n_h \log \theta + n_t \log(1-\theta) \right]\end{aligned}$$

- First order condition (equating the derivative to zero):

$$\frac{n_h}{\theta} - \frac{n_t}{1-\theta} = 0 \iff \theta = \frac{n_h}{n_h + n_t}$$

$\hat{\theta}_{\text{MLE}}$ is the empirical fraction of heads.

Table of Contents

Bayesian Statistics

- Baysian statistics introduces a crucial new ingredient: the **prior distribution.**
- A **prior distribution** $p(\theta)$ is a distribution on the parameter space Θ.
- The prior reflects our belief about θ, **before seeing any data**.

A Bayesian Model

- A [parametric] Bayesian model consists of two pieces:
  1. A parametric family of densities
  $$\{p(\mathcal{D} \mid \theta) \mid \theta \in \Theta\}.$$
  2. A **prior distribution** $p(\theta)$ on parameter space Θ.
- Putting the pieces together, we get a joint density on θ and $\mathcal{D}$:
$$p(\mathcal{D}, \theta) = p(\mathcal{D} \mid \theta)p(\theta).$$

The Posterior Distribution

- The **posterior distribution** for θ is $p(\theta \mid \mathcal{D})$.
- Whereas the prior represents belief about θ before observing data $\mathcal{D}$,
- The posterior represents the **rationally updated belief** about θ, after seeing $\mathcal{D}$.

Expressing the Posterior Distribution

- By Bayes rule, can write the posterior distribution as

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

- Let's consider both sides as functions of θ, for fixed $\mathcal{D}$.
- Then both sides are densities on Θ and we can write

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

- Where $\propto$ means we've dropped factors that are independent of θ.

Coin Flipping: Bayesian Model

- Recall that we have a parametric family of mass functions:

$$p(\text{Heads} \mid \theta) = \theta,$$

  for $\theta \in \Theta = (0, 1)$.
- We need a prior distribution $p(\theta)$ on $\Theta = (0, 1)$.
- One convenient choice would be a distribution from the Beta family

Coin Flipping: Beta Prior

- **Prior:**

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

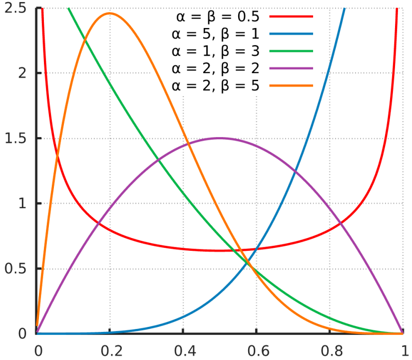


Figure by Horas based on the work of Krishnavedala (Own work) [Public domain], via Wikimedia Commons http://commons.wikimedia.org/wiki/File:Beta_distribution_pdf.svg.

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:**

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

- **Mode of Beta distribution:**

$$\arg\max_{\theta} p(\theta) = \frac{h-1}{h+t-2}$$

for $h,t>1$.

Coin Flipping: Posterior

- **Prior:**

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

- **Likelihood function**

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

- **Posterior density:**

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

The Posterior is in the Beta Family!

- **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.

Sidebar: Conjugate Priors

- In this case, the posterior is in the same distribution family as the prior.
- Let π be a family of prior distributions on Θ.
- Let P parametric family of distributions with parameter space Θ.

Definition

A family of distributions π **is conjugate to** parametric model P if for any prior in π, the posterior is always in π.

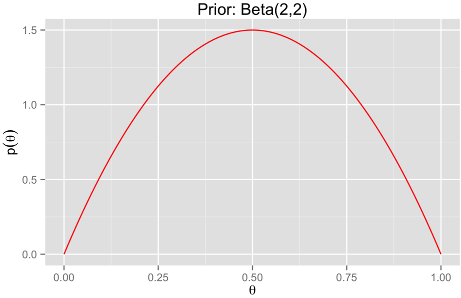
- The beta family is conjugate to the coin-flipping (i.e. Bernoulli) model.

Coin Flipping: Concrete Example

- Suppose we have a coin, possibly biased (**parametric probability model**):

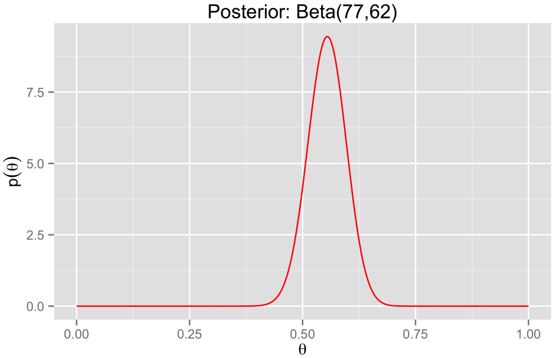
$$p(\text{Heads} \mid \theta) = \theta.$$

- **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

- We have the posterior distribution $\theta \mid \mathcal{D}$.
- What if someone asks us for a point estimate $\hat{\theta}$ for θ?
- 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: display uncertainty estimates to our client
- Extract a **credible set** for θ (a Bayesian confidence interval).
  - e.g. Interval $[a, b]$ is a 95% **credible set** if

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

- Select a point estimate using **Bayesian decision theory**:
  - Choose a loss function.
  - Find action **minimizing expected risk w.r.t. posterior**

Table of Contents

Bayesian Decision Theory

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

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

  - 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 $p(y \mid \theta)$, for unknown $\theta \in \Theta$.
  - We want to produce a **point estimate** for θ.
- Choose:
  - **Prior** $p(\theta)$ on $\Theta = \mathrm{R}$.
  - **Loss** $\ell(\hat{\theta}, \theta)$
- Find **action** $\hat{\theta} \in \Theta$ that minimizes the **posterior risk:**

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

Important Cases

- Squared Loss : $\ell(\hat{\theta}, \theta) = \left(\theta - \hat{\theta}\right)^2 \quad \Rightarrow$ posterior mean
- Zero-one Loss: $\ell(\theta, \hat{\theta}) = 1(\theta \neq \hat{\theta}) \quad \Rightarrow$ posterior mode
- Absolute Loss : $\ell(\hat{\theta}, \theta) = \left|\theta - \hat{\theta}\right| \quad \Rightarrow$ posterior median

Bayesian Point Estimation: Square Loss

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

$$r(\hat{\theta}) = \int \left(\theta - \hat{\theta}\right)^2 p(\theta \mid \mathcal{D})\, d\theta.$$

- Differentiate:

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

Bayesian Point Estimation: Square Loss

- Derivative of posterior risk is

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

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

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

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

Table of Contents

Recap and Interpretation

- The prior represents belief about θ before observing data $\mathcal{D}$.
- The posterior represents **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 justifying an estimator.
    - Only choices are
        - **family of distributions**, indexed by Θ, and
        - **prior distribution** on Θ
    - For decision making, we need a **loss function**.

Table of Contents

Conditional Probability Modeling

- **Input space** $\mathcal{X}$
- **Outcome space** $\mathcal{Y}$
- **Action space** $\mathcal{A} = \{p(y) \mid p$ is a probability distribution on $\mathcal{Y}\}$.
- **Hypothesis space** $\mathcal{F}$ contains prediction functions $f : \mathcal{X} \to \mathcal{A}$.
- **Prediction function** $f \in \mathcal{F}$ takes input $x \in \mathcal{X}$ and produces a **distribution** on $\mathcal{Y}$
- A **parametric family of conditional densities** is a set

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

  - where $p(y \mid x, \theta)$ is a density on **outcome space** $\mathcal{Y}$ for each x in **input space** $\mathcal{X}$, and
  - θ is a **parameter** in a [finite dimensional] **parameter space** Θ**.**

- This is the common starting point for either classical or Bayesian regression.

Classical treatment: Likelihood Function

- **Data:** $\mathcal{D} = (y_1, \ldots, , y_n)$
- The probability density for our data $\mathcal{D}$ is

$$p(\mathcal{D} \mid x_1, \ldots, x_n, \theta) \quad = \quad \prod_{i=1}^{n} p(y_i \mid x_i, \theta).$$

- For fixed $\mathcal{D}$, the function $\theta \mapsto p(\mathcal{D} \mid x, \theta)$ is the **likelihood function**:

$$L_{\mathcal{D}}(\theta) = p(\mathcal{D} \mid x, \theta),$$

where $x = (x_1, \ldots, x_n)$.

Maximum Likelihood Estimator

- The **maximum likelihood estimator (MLE)** for θ in the family $\{p(y \mid x, \theta) \mid \theta \in \Theta\}$ is

$$\hat{\theta}_{\text{MLE}} = \underset{\theta \in \Theta}{\arg\max}\, L_{\mathcal{D}}(\theta).$$

- MLE corresponds to ERM, if we set the loss to be the negative log-likelihood.
- The corresponding prediction function is

$$\hat{f}(x) = p(y \mid x, \hat{\theta}_{\text{MLE}}).$$

Table of Contents

Bayesian Conditional Models

- Input space $\mathcal{X} = \mathbf{R}^d$ Outcome space $\mathcal{Y} = \mathbf{R}$
- The Bayesian conditional model has two components:
  - A **parametric family of conditional densities**:

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

  - A **prior distribution** $p(\theta)$ on $\theta \in \Theta$.

The Posterior Distribution

- The **prior distribution** $p(\theta)$ represents our beliefs about θ before seeing $\mathcal{D}$.
- The **posterior distribution** for θ is

$$
\begin{aligned}
p(\theta \mid \mathcal{D}, x) &\propto p(\mathcal{D} \mid \theta, x) p(\theta) \\
&= \underbrace{L_{\mathcal{D}}(\theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}
\end{aligned}
$$

- Posterior represents the **rationally updated beliefs** after seeing $\mathcal{D}$.
- Each θ corresponds to a prediction function,
  - i.e. the conditional distribution function $p(y \mid x, \theta)$.

Point Estimates of Parameter

- What if we want point estimates of θ?
- We can use **Bayesian decision theory** to derive point estimates.
- We may want to use
  - $\hat{\theta} = \mathbb{E}[\theta \mid \mathcal{D}, x]$ (the posterior mean estimate)
  - $\hat{\theta} = \text{median}[\theta \mid \mathcal{D}, x]$
  - $\hat{\theta} = \arg\max_{\theta \in \Theta} p(\theta \mid \mathcal{D}, x)$ (the MAP estimate)
- depending on our loss function.

Back to the basic question - Bayesian Prediction Function

- Find a function takes input $x \in \mathcal{X}$ and produces a **distribution** on $\mathcal{Y}$
- In the frequentist approach:
  - Choose family of conditional probability densities (hypothesis space).
  - Select one conditional probability from family, e.g. using MLE.
- In the Bayesian setting:
  - We choose a parametric family of conditional densities

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

  - and a prior distribution $p(\theta)$ on this set.
- Having set our Bayesian model, how do we predict a distribution on y for input x?
- We don't need to make a discrete selection from the hypothesis space: we **maintain uncertainty**.

The Prior Predictive Distribution

- Suppose we have not yet observed any data.
- In the Bayesian setting, we can still produce a prediction function.
- The **prior predictive distribution** is given by

$$x \mapsto p(y \mid x) = \int p(y \mid x; \theta) p(\theta) \, d\theta.$$

- This is an average of all conditional densities in our family, weighted by the prior.

The Posterior Predictive Distribution

- Suppose we've already seen data $\mathcal{D}$.
- The **posterior predictive distribution** is given by

$$x \mapsto p(y \mid x, \mathcal{D}) = \int p(y \mid x; \theta) p(\theta \mid \mathcal{D})\, d\theta.$$

- This is an average of all conditional densities in our family, weighted by the posterior.

Comparison to Frequentist Approach

- In Bayesian statistics we have two distributions on Θ:
  - the prior distribution $p(\theta)$
  - the posterior distribution $p(\theta \mid \mathcal{D})$.
- These distributions over parameters correspond to distributions on the hypothesis space:

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

- In the frequentist approach, we choose $\hat{\theta} \in \Theta$, and predict

$$p(y \mid x, \hat{\theta}(\mathcal{D})).$$

- In the Bayesian approach, we integrate out over Θ w.r.t. $p(\theta \mid \mathcal{D})$ and predict with

$$p(y \mid x, \mathcal{D}) = \int p(y \mid x; \theta) p(\theta \mid \mathcal{D})\, d\theta$$

What if we don't want a full distribution on y?

- Once we have a predictive distribution $p(y \mid x, \mathcal{D})$,
  - we can easily generate single point predictions.
- $x \mapsto \mathbb{E}[y \mid x, \mathcal{D}]$, to minimize expected square error.
- $x \mapsto \text{median}[y \mid x, \mathcal{D}]$, to minimize expected absolute error
- $x \mapsto \arg\max_{y \in \mathcal{Y}} p(y \mid x, \mathcal{D})$, to minimize expected 0/1 loss
- Each of these can be derived from $p(y \mid x, \mathcal{D})$.

Table of Contents

Example in 1-Dimension: Setup

- Input space $\mathcal{X} = [-1, 1]$ Output space $\mathcal{Y} = \mathbf{R}$
- Given x, the world generates y as

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

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

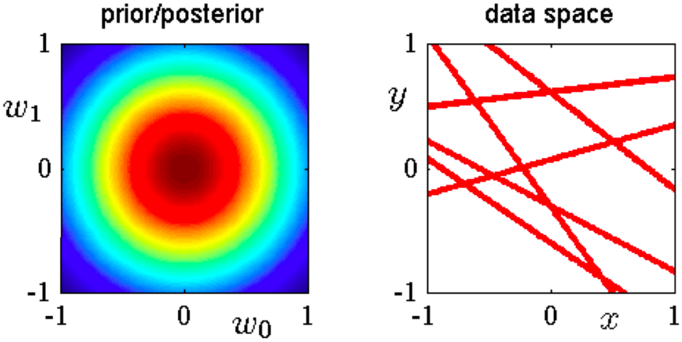
- Written another way, the **conditional probability model** is

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

- What's the parameter space? $\mathbf{R}^2$.
- **Prior distribution:** $w = (w_0, w_1) \sim \mathcal{N}\left(0, \frac{1}{2} I\right)$

Example in 1-Dimension: Prior Situation

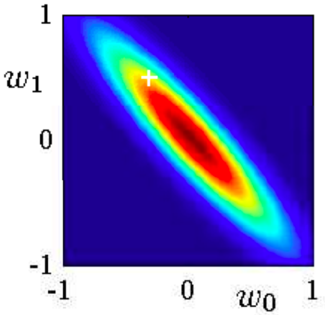
- **Prior distribution:** $w = (w_0, w_1) \sim \mathcal{N}\left(0, \frac{1}{2}I\right)$ (Illustrated on left)

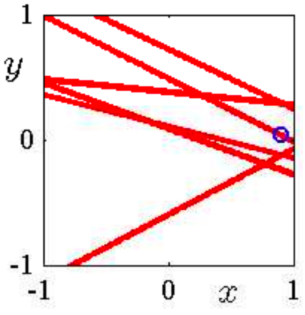


- On right, $y(x) = \mathbb{E}[y \mid x, w] = w_0 + w_1 x$, for randomly chosen $w \sim p(w) = \mathcal{N}\left(0, \frac{1}{2}I\right)$.

Bishop's PRML Fig 3.7

Example in 1-Dimension: 1 Observation

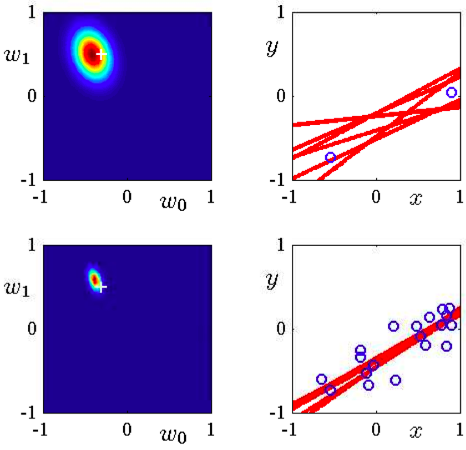




- On left: posterior distribution; white cross indicates true parameters
- On right:
  - blue circle indicates the training observation
  - red lines, $y(x) = \mathbb{E}[y \mid x, w] = w_0 + w_1 x$, for randomly chosen $w \sim p(w|\mathcal{D})$ (posterior)

Bishop's PRML Fig 3.7

Example in 1-Dimension: 2 and 20 Observations



Bishop's PRML Fig 3.7

Table of Contents

Closed Form for Posterior

- Model:

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

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

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

- **Posterior Variance** Σ_P gives us a natural **uncertainty measure.**

Closed Form for Posterior

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

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

- If we want point estimates of w, **MAP estimator** and the **posterior mean** are given by

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

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

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

which is of course the ridge regression solution.