Bayesian Methods

He He
(adapted from David Rosenberg's slides)

CDS, NYU

March 31, 2019

Contents

Logistics

- Deliverable
  - Homework 4 due today
  - Homework 5 released (due on April 15)
  - Project proposal due tomorrow
    - Make sure each group has a corresponding member
- Codalab
  - Optional but bonus points if you provide a worksheet for reproducibility
  - Tutorial this Thursday during the instructor's office hour

Introduction

Recap: typical steps in data science problems

Many problem domains can be formalized as follows:

1. Observe input x.
2. Take action a.
3. Observe outcome y.
4. Evaluate action in relation to the outcome (via a loss function $\ell(a, y)$)

The Three Spaces:

- Input space: $\mathcal{X}$
- Action space: $\mathcal{A}$
- Outcome space: $\mathcal{Y}$

Some Formalization

The Spaces

- $\mathcal{X}$: input space
- $\mathcal{Y}$: outcome space
- $\mathcal{A}$: action space

Prediction Function (or "decision function")

A **prediction function** (or **decision function**) gets input $x \in \mathcal{X}$ and produces an action $a \in \mathcal{A}$:

$$\begin{aligned} \delta : \quad \mathcal{X} &\to \mathcal{A} \\ x &\mapsto f(x) \end{aligned}$$

Loss Function

A **loss function** evaluates an action in the context of the outcome y.

$$\begin{aligned} \ell : \quad \mathcal{A} \times \mathcal{Y} &\to \mathbf{R} \\ (a, y) &\mapsto \ell(a, y) \end{aligned}$$

Statistical inference

- Observe data $\mathcal{D} = \{y_1, \ldots, y_N\}$
- Assume that data is generated by a family of parametric distributions

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

  - 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.
- Assume that data is drawn i.i.d. from $p(y \mid \theta)$.
- The decision-making problem: Infer properties of $p(y \mid \theta)$ given some observed data

Today's lecture

- How to make decisions given unknown nature/world and limited data?
  - The frequentist approach
  - The Bayesian approach
- Apply the Bayesian approach to conditional models (classification)
  - Learning and prediction

Frequentist Decision Theory

Frequentist or "Classical" Statistics

Key idea:

- There exists a true but unknown parameter θ^*.
- We can obtain its estimate $\hat{\theta}$ from a sample $\mathcal{D} \sim p(\mathcal{D} \mid \theta^*)$ using some **point estimator** δ.
  - In general, $\delta : \mathcal{X} \to \mathcal{A}$ is a decision procedure based on data.

Task: estimate θ given i.i.d. samples from $p(y \mid \theta)$ where $\theta \in \Theta$.

How do we choose the best estimator?

$$\textbf{Frequentist risk:} \quad R(\theta^*, \delta) = \mathbb{E}_{p(\mathcal{D}\mid\theta^*)} L(\theta^*, \delta(\mathcal{D})) \tag{1}$$

But we don't know θ^*...

Desirable Properties of Estimators

Heuristics for selecting a good estimator:

- **Consistent**: As data size $N \to \infty$, we get $\hat{\theta} \to \theta^*$.
  - What assumptions are we making here?
- **Unbiased**: our estimate is correct in expectation.

$$\bar{\theta} \stackrel{\text{def}}{=} \mathbb{E}_{p(\mathcal{D}|\theta^*)}\left[\hat{\theta}\right] = \theta^* \tag{2}$$

$$\text{bias}(\hat{\theta}) = \bar{\theta} - \theta^* \tag{3}$$

- **Minimum variance**:

$$\text{var}(\hat{\theta}) = \mathbb{E}_{p(\mathcal{D}|\theta^*)}\left[\left(\hat{\theta} - \bar{\theta}\right)^2\right] \tag{4}$$

Desirable Properties of Estimators

Heuristics for selecting a good estimator:

- **Consistent**: As data size $N \to \infty$, we get $\hat{\theta} \to \theta^*$.
  - What assumptions are we making here?
- **Unbiased**: our estimate is correct in expectation.

$$\bar{\theta} \stackrel{\text{def}}{=} \mathbb{E}_{p(\mathcal{D}|\theta^*)}\left[\hat{\theta}\right] = \theta^* \tag{2}$$

$$\text{bias}(\hat{\theta}) = \bar{\theta} - \theta^* \tag{3}$$

- **Minimum variance**:

$$\text{var}(\hat{\theta}) = \mathbb{E}_{p(\mathcal{D}|\theta^*)}\left[\left(\hat{\theta} - \bar{\theta}\right)^2\right] \tag{4}$$

Observed data is actually generated from θ^*

The bias-variance tradeoff

Do we always want an unbiased estimator?

Let's decompose the square loss. (expectations are over $p(\mathcal{D} \mid \theta^*)$)

$$
\begin{aligned}
\mathbb{E}\left[\left(\hat{\theta}-\theta^*\right)^2\right] &= \mathbb{E}\left[\left(\hat{\theta}-\bar{\theta}+\bar{\theta}-\theta^*\right)^2\right] && (5)\\
&= \mathbb{E}\left[\left(\hat{\theta}-\bar{\theta}\right)^2\right] + 2(\bar{\theta}-\theta^*)\underbrace{\mathbb{E}\left[\left(\hat{\theta}-\bar{\theta}\right)\right]}_{=0 \text{ because } \bar{\theta} \stackrel{\text{def}}{=} \mathbb{E}\left[\hat{\theta}\right]} + \mathbb{E}\left[(\bar{\theta}-\theta^*)^2\right] && (6)\\
&= \mathbb{E}\left[\left(\hat{\theta}-\bar{\theta}\right)^2\right] + (\bar{\theta}-\theta^*)^2 && (7)\\
&= \text{var}(\hat{\theta}) + \text{bias}^2(\hat{\theta}) && (8)
\end{aligned}
$$

Example: ridge regression

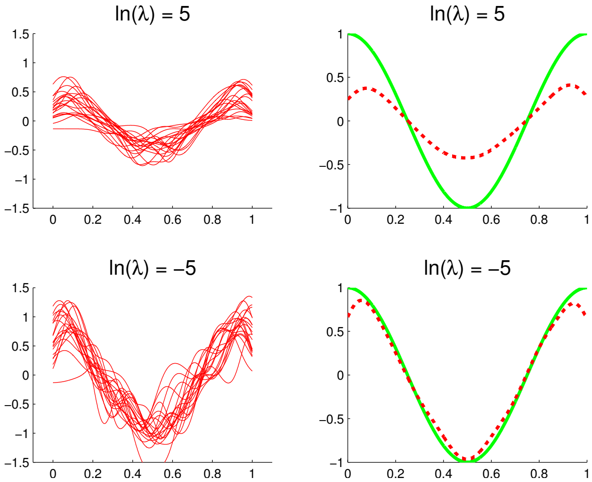


Figure 6.5 from "Machine Learning: a Probabilistic Perspective", K. Murphy.

Maximum Likelihood Estimation

Definition

The **maximum likelihood estimator (MLE)** for θ in the model $\{p(y \mid \theta) : \theta \in \Theta\}$ is

$$\hat{\theta}_{\text{MLE}} = \underset{\theta \in \Theta}{\arg\max}\, L_{\mathcal{D}}(\theta), \tag{9}$$

$$\text{where } L_{\mathcal{D}}(\theta) \stackrel{\text{def}}{=} p(\mathcal{D} \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta) \tag{10}$$

- MLE is consistent but can be biased.
- **Method of moments** is another general approach one learns about in statistics.

Example: Coin Flipping

Task: mode a biased coin.

- Parametric family of mass functions:

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

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

- Note that every $\theta \in \Theta$ gives us a different probability model for a coin.

Coin Flipping: Likelihood function

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

$$L_{\mathcal{D}}(\theta) = p(\mathcal{D} \mid \theta) = \theta^{n_h}(1-\theta)^{n_t} \tag{11}$$

Coin Flipping: MLE

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

$$\begin{aligned} \hat{\theta}_{\text{MLE}} &= \underset{\theta \in \Theta}{\arg\max} \log L_{\mathcal{D}}(\theta) && (12) \\ &= \underset{\theta \in \Theta}{\arg\max} \left[n_h \log \theta + n_t \log(1-\theta) \right] && (13) \end{aligned}$$

- First order condition:

$$\begin{aligned} \frac{\partial}{\partial \theta} \ell = \frac{n_h}{\theta} - \frac{n_t}{1-\theta} &= 0 && (14) \\ \iff \theta &= \frac{n_h}{n_h + n_t}. && (15) \end{aligned}$$

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

Review

Challenges in statistical inference:

- Unknown data generating process defined by θ
- Cannot observe all data
- Want to infer properties of θ (and make decisions/predictions)

Frequentist approach:

- Point estimator based on a data sample
- Compare estimators by expected loss over all possible data samples—impossible
- Other metrics: consistency, unbiasedness, variance etc.
- A common estimator: MLE

Next, the Bayesian approach.

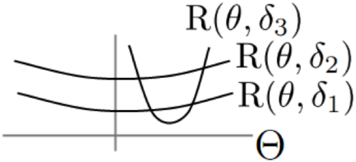
Bayesian Decision Theory

Bayesian twist of the frequentist risk

Task Design a measure to evaluate some estimator δ.

Problem cannot compute the risk without knowing θ^*.

$$R(\theta^*, \delta) = \mathbb{E}_{p(\mathcal{D}|\theta^*)} L(\theta^*, \delta(\mathcal{D})) \tag{16}$$



Solution introduce the prior $p(\theta^*)$.

$$\textbf{Bayes risk}: R_B(\delta) = \int R(\theta^*, \delta) p(\theta^*) d\theta^* \tag{17}$$

Note Bayes risk is a frequentist concept because it still averages over the data $p(\mathcal{D} \mid \theta^*)$.

The Bayesian approach

Key idea:

- The true θ is never known but we have **belief** about it (no more θ^*)
- As we observe more data, we can update our beliefs (no expectation over unseen data)

Key concepts:

Prior $p(\theta)$, our belief before seeing any data.

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

Marginal likelihood $p(\mathcal{D}) = \int p(\mathcal{D} \mid \theta) p(\theta) d\theta$ (also called evidence)

Posterior probability $p(\theta \mid \mathcal{D})$, our updated belief after seeing $\mathcal{D}$.

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

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 independent of θ.

Posterior risk

Bayesian interpretation of the risk: posterior expected loss.

$$\textbf{posterior risk: } r(a \mid \mathcal{D}, p(\theta)) \stackrel{\text{def}}{=} \mathbb{E}_{p(\theta \mid \mathcal{D})}\left[L(\theta, a)\right] \qquad \text{where } a = \delta(\mathcal{D}) \tag{18}$$

- Conditioned on observed data and the prior, which are known.
- Average over the posterior distribution of θ.

How to make decisions?

$$\textbf{Bayes action: } \delta^*(\mathcal{D}) \stackrel{\text{def}}{=} \underset{a \in \mathcal{A}}{\arg\min}\, \mathbb{E}_{p(\theta \mid \mathcal{D})}\left[L(\theta, a)\right] \tag{19}$$

- No need to choose an estimator.
- What might be the practical issue here?

Posterior risk

Bayesian interpretation of the risk: posterior expected loss.

$$\textbf{posterior risk: } r(a \mid \mathcal{D}, p(\theta)) \stackrel{\text{def}}{=} \mathbb{E}_{p(\theta \mid \mathcal{D})}[L(\theta, a)] \qquad \text{where } a = \delta(\mathcal{D}) \tag{18}$$

- Conditioned on observed data and the prior, which are known.
- Average over the posterior distribution of θ.

How to make decisions?

$$\textbf{Bayes action: } \delta^*(\mathcal{D}) \stackrel{\text{def}}{=} \underset{a \in \mathcal{A}}{\arg\min}\, \mathbb{E}_{p(\theta \mid \mathcal{D})}[L(\theta, a)] \tag{19}$$

- No need to choose an estimator.
- What might be the practical issue here?

1. How is it different from the frequentist risk?
2. compute the expectation

Coin Flipping: Bayesian Model

- Parametric family of mass functions:

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

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

- Need a prior distribution $p(\theta)$ on $\Theta = (0,1)$.

- Likelihood $p(x \mid \theta)$ is Bernoulli.
- A distribution from the Beta family will do the trick...

Coin Flipping: Beta Prior

$$\theta \sim \text{Beta}(\alpha, \beta) \tag{20}$$

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{21}$$

$$\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta} \tag{22}$$

Think of α and β as our initial counts of head (h) and tails (t) before seeing any data.

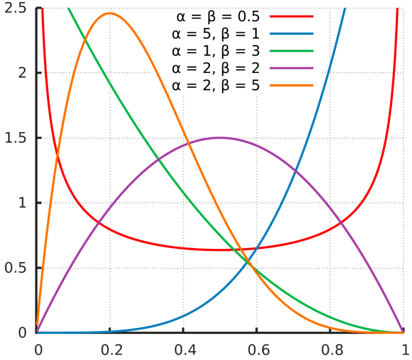


Figure by Horas based on the work of Krishnavedala (Own work) [Public domain], via Wikimedia Commons http://commons.wikimedia.org/wiki/File:Beta_distribution_pdf.svg.

Coin Flipping: Posterior

- **Prior:**

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

- **Likelihood function**

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

- **Posterior density:**

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

What is the posterior distribution?

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.

Posterior is Beta

- **Prior:**

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

- **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.

This leads us back to the previous question—why use a Beta prior?

Conjugate Priors

Interesting that posterior is in the same distribution family as prior.

Definition

A family of priors π **is conjugate to** a parametric model P (the likelihood) if the posterior is in the same family π.

Examples:

- The beta family is conjugate to the coin-flipping (i.e. Bernoulli) model.
- The family of all probability distributions is conjugate to any parametric model. [Trvially]

Why use conjugate priors? Mainly for computational convenience.

Compute the posterior in Coin Flipping

Likelihood $p(\text{Heads} \mid \theta) = \theta$ for $\theta \in \Theta = [0, 1]$.

Prior $\theta \sim \text{Beta}(2, 2)$.

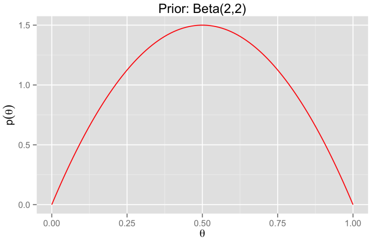
Data $\mathcal{D} = \{H, H, T, \ldots, T\}$, 75 heads, 60 tails

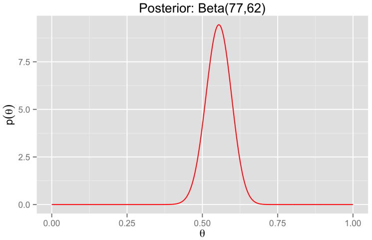
Posterior $\theta \mid \mathcal{D} \sim \text{Beta}(77, 62)$

MLE $\hat{\theta}_{\text{MLE}} = \frac{75}{75+60} \approx 0.556$

- When might the MLE estimate be bad?

Given the posterior, what would be a good estimate of the value θ?

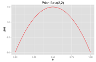


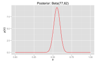


Compute the posterior in Coin Flipping

- **Likelihood** $p(\text{Heads} \mid \theta) = \theta$ for $\theta \in \Theta = [0,1]$.
- **Prior** $\theta \sim \text{Beta}(2,2)$.
- **Data** $\mathcal{D} = \{H, H, T, \ldots, T\}$, 75 heads, 60 tails
- **Posterior** $\theta \mid \mathcal{D} \sim \text{Beta}(77, 62)$
- **MLE** $\hat{\theta}_{\text{MLE}} = \frac{75}{75+60} \approx 0.556$
  - When might the MLE estimate be bad?

Given the posterior, what would be a good estimate of the value θ?





few observationse.g. posterior mean

Bayesian point estimation

Setup:

- Data $\mathcal{D}$ generated by $p(y \mid \theta)$, for unknown $\theta \in \Theta$.
- Want to produce a point estimate for θ.

Approach:

1. Choose a loss function, e.g., square loss $L(\theta, \hat{\theta}) = \left(\theta - \hat{\theta}\right)^2$.
2. Find an action minimizing the expected risk w.r.t. posterior—Bayes action.

Bayesian Point Estimation: Square Loss

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

$$r(\hat{\theta}) = \int \left(\theta - \hat{\theta}\right)^2 p(\theta \mid \mathcal{D})\, d\theta. \tag{23}$$

- Differentiate:

$$\frac{dr(\hat{\theta})}{d\hat{\theta}} = -\int 2\left(\theta - \hat{\theta}\right) p(\theta \mid \mathcal{D})\, d\theta \tag{24}$$

$$= -2\int \theta p(\theta \mid \mathcal{D})\, d\theta + 2\hat{\theta} \underbrace{\int p(\theta \mid \mathcal{D})\, d\theta}_{=1} \tag{25}$$

$$= -2\int \theta p(\theta \mid \mathcal{D})\, d\theta + 2\hat{\theta} \tag{26}$$

- Set to zero:

$$\hat{\theta} = \int \theta p(\theta \mid \mathcal{D})\, d\theta = \mathbb{E}\left[\theta \mid \mathcal{D}\right] \qquad \text{posterior mean} \tag{27}$$

Bayesian Point Estimation: Absolute Loss

- Posterior risk:

$$r(\hat{\theta}) = \int \left|\theta - \hat{\theta}\right| p(\theta \mid \mathcal{D})\, d\theta. \tag{28}$$

$$= \int_{-\infty}^{\hat{\theta}} \left(\hat{\theta} - \theta\right) p(\theta \mid \mathcal{D})\, d\theta + \int_{\hat{\theta}}^{\infty} \left(\theta - \hat{\theta}\right) p(\theta \mid \mathcal{D})\, d\theta \tag{29}$$

- Differentiate:

$$\frac{dr(\hat{\theta})}{d\hat{\theta}} = \int_{-\infty}^{\hat{\theta}} p(\theta \mid \mathcal{D})\, d\theta - \int_{\hat{\theta}}^{\infty} p(\theta \mid \mathcal{D})\, d\theta \tag{30}$$

- Set to zero:

$$\int_{-\infty}^{\hat{\theta}} p(\theta \mid \mathcal{D})\, d\theta = \int_{\hat{\theta}}^{\infty} p(\theta \mid \mathcal{D})\, d\theta \qquad \text{and they sum to one} \tag{31}$$

$$\implies \hat{\theta} \text{ split the area under the curve evenly: posterior median} \tag{32}$$

Bayesian Point Estimation: Zero-One Loss

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

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

- **Bayes action** is

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

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

Review: the Bayesian method

1. **Define the model**:
   - Choose a parametric family of densities—likelihood:
   $$\{p(\mathcal{D} \mid \theta) \mid \theta \in \Theta\}.$$
   - Choose a distribution $p(\theta)$ on Θ—prior distribution.
2. After observing data $\mathcal{D}$, compute the posterior distribution $p(\theta \mid \mathcal{D})$.
3. Choose action based on $p(\theta \mid \mathcal{D})$ and the loss function.

Frequentist vs Bayesian

	Frequentist	Bayesian
Evaluate a decision	$L(\theta, \delta(\cdot))$	$L(\theta, \delta(\cdot))$
Handle unknown state of nature (θ)	θ^*	θ is a variable—prior, posterior
Make decisions	average over (observed and unobserved) data	average over θ
Topics of interests	properties of an estimator (e.g., consistent, unbiased)	compute various quantities, e.g., posterior, marginal etc.
History	dominated during the 20th century	dominated before the 20th century

Bayesian Conditional Models

Learning as density estimation

Setup

- Observe data $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N}$ assuming $x^{(n)}$'s are fixed.
- Choose a family of parametric distributions:

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

Learning

- Maximum likelihood estimation:

$$\hat{\theta}_{\text{MLE}} = \underset{\theta \in \Theta}{\arg\max}\, L_{\mathcal{D}}(\theta) = \underset{\theta \in \Theta}{\arg\max}\, p(\mathcal{D} \mid \theta, x) \tag{33}$$

- Assume $y^{(n)}$'s are independent conditioned on $x^{(n)}$.
- Exercise: MLE corresponds to ERM with negative log-likelihood loss.

Prediction

$$p(y \mid x, \hat{\theta}_{\text{MLE}}) \tag{34}$$

Example: Gaussian linear regression

Model

$$p(y \mid x, \theta) = \mathcal{N}(\theta^T x, \sigma^2) \qquad \text{Assuming known } \sigma^2. \tag{35}$$

Likelihood

$$L_{\mathcal{D}}(\theta) = \prod_{n=1}^{N} p(y^{(n)} \mid x^{(n)}, \theta) \tag{36}$$

$$= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(n)} - \theta^T x^{(n)}\right)^2}{2\sigma^2}\right) \tag{37}$$

Solution

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \mathbf{R}^d} L_{\mathcal{D}}(\theta) \tag{38}$$

$$= \arg\max_{\theta \in \mathbf{R}^d} \sum_{n=1}^{N} \left(y^{(n)} - \theta^T x^{(n)}\right)^2 \qquad \text{squared loss} \tag{39}$$

- We want small weights to avoid overfitting. What would be a good prior?

$$\theta \sim \mathcal{N}\left(0, \tau^2 I_d\right) \qquad \text{Why Gaussian?} \tag{40}$$

- Posterior distribution is also a Gaussian distribution:

$$p(\theta \mid \mathcal{D}) \propto \mathcal{N}(0, \tau^2 I_d)\mathcal{N}(X\theta, \sigma^2 I_N) \tag{41}$$

$$= \mathcal{N}(\mu_P, \Sigma_P) \tag{42}$$

$$\mu_P = \left(X^T X + \frac{\sigma^2}{\tau^2} I_d\right)^{-1} X^T y \tag{43}$$

$$\Sigma_P = \left(\sigma^{-2} X^T X + \tau^{-2} I_d\right)^{-1} \tag{44}$$

- See Rosenberg's notes on multivariate Gaussian.

MAP (instead of MLE)

- Instead of maximizing the likelihood, let's maximize the posterior distribution to incorporate the prior.

$$p(\theta \mid \mathcal{D}) \propto \underbrace{\exp\left(-\frac{1}{2\tau^2}\|\theta\|^2\right)}_{\text{prior}} \underbrace{\prod_{i=1}^{n} \exp\left(-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}\right)}_{\text{likelihood}} \tag{45}$$

- To find MAP, sufficient to minimize the negative log posterior (Exercise):

$$\hat{\theta}_{\text{MAP}} = \operatorname*{arg\,min}_{\theta \in \mathbf{R}^d} \left[-\log p(\theta \mid \mathcal{D})\right] \tag{46}$$

$$= \operatorname*{arg\,min}_{\theta \in \mathbf{R}^d} \underbrace{\sum_{i=1}^{n} (y_i - \theta^T x_i)^2}_{\text{log-likelihood}} + \underbrace{\lambda \|\theta\|^2}_{\text{log-prior}} \qquad \lambda \stackrel{\text{def}}{=} \frac{\sigma^2}{\tau^2} \tag{47}$$

- How does the prior control the regularization strength?

The Bayesian approach

- In Bayesian setting, **there is no selection** from hypothesis space, e.g., $\hat{\theta}_{\text{MLE}}, \hat{\theta}_{\text{MAP}}$.
- We chose 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, there are no more decisions to make – just computation...
  - posterior distribution
  - predictive distribution

Bayesian conditional models

- 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 updated beliefs after seeing $\mathcal{D}$.

Bayesian linear regression

Let's derive ridge regression from a Bayesian perspective.

- Gaussian prior: $\theta \sim \mathcal{N}(0, \Sigma_0)$.
- Posterior distribution is also Gaussian:

$$
\begin{aligned}
\theta \mid \mathcal{D} &\sim \mathcal{N}(\mu_P, \Sigma_P) &(48)\\
\mu_\mathrm{P} &= \left(X^TX + \sigma^2\Sigma_0^{-1}\right)^{-1} X^T y &(49)\\
\Sigma_\mathrm{P} &= \left(\sigma^{-2}X^TX + \Sigma_0^{-1}\right)^{-1} &(50)
\end{aligned}
$$

- What are reasonable point estimates of θ? Posterior mode (MAP) and posterior mean:

$$\hat{\theta} = \mu_P = \left(X^TX + \sigma^2\Sigma_0^{-1}\right)^{-1} X^T y \qquad \text{familiar?} \tag{51}$$

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

$$\hat{w} = \mu_P = \left(X^TX + \lambda I\right)^{-1} X^T y, \qquad \text{ridge regression.} \tag{52}$$

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

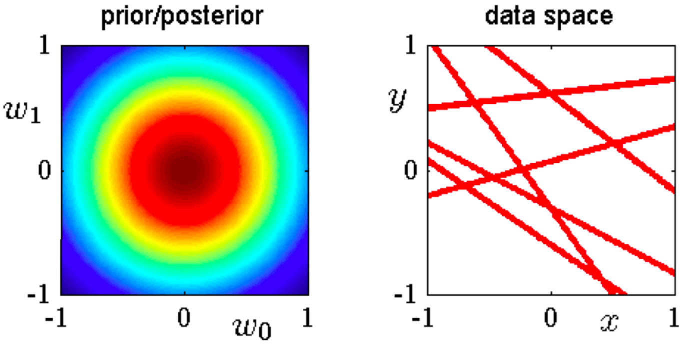
- Written another way, the **conditional probability model** is

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

- 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

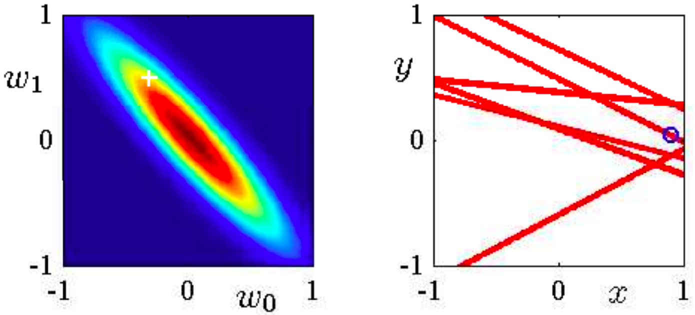
- **Prior distribution:** $w = (w_0, w_1) \sim \mathcal{N}\left(0, \frac{1}{2}I\right)$



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

Bishop's PRML Fig 3.7

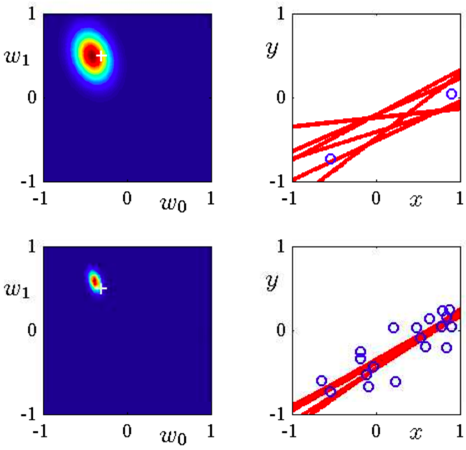
Example in 1-Dimension: 1 Observation



- On left: posterior distribution; white '+' indicates true parameters
- On right: 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

Review

- Task: find a function in a hypothesis space that map x to a distribution of y:

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

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

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

- In Bayesian statistics we have two distributions on Θ:
  - the prior distribution $p(\theta)$
  - the posterior distribution $p(\theta \mid \mathcal{D})$.
- Next, prediction by integrating over Θ w.r.t. $p(\theta \mid \mathcal{D})$.

The Predictive Distribution

- Without any data, the **prior predictive distribution** is given by

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

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

- Once we see data $\mathcal{D}$, the **posterior predictive distribution** is given by

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

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

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

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.
- $x \mapsto \operatorname{median}[y \mid x, \mathcal{D}]$, 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 $p(y \mid x, \mathcal{D})$.

Remember when we talked about Bayesian point estimation, we can derive the Bayes action given a posterior distribution and a loss function.

Bayesian linear regression: Predictive Distribution

Let's go back to Gaussian linear regression:

$$\theta \sim \mathcal{N}(0, \Sigma_0) \qquad \text{prior} \tag{53}$$

$$y^{(n)} \mid x^{(n)}, \theta \sim \mathcal{N}(\theta^T x^{(n)}, \sigma^2) \qquad \text{likelihood} \tag{54}$$

Predictive Distribution

$$p(y_{\text{new}} \mid x_{\text{new}}, \mathcal{D}) = \int p(y_{\text{new}} \mid x_{\text{new}}, \theta) p(\theta \mid \mathcal{D})\, d\theta \tag{55}$$

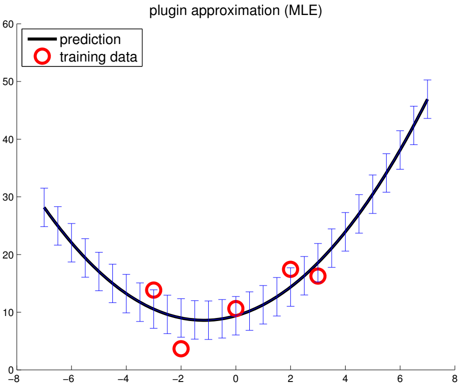
$$= \mathcal{N}(\eta_{\text{new}}, \sigma^2_{\text{new}}) \qquad \text{also a Gaussian} \tag{56}$$

$$\eta_{\text{new}} = \mu_P^T x_{\text{new}} \qquad \text{MAP prediction} \tag{57}$$

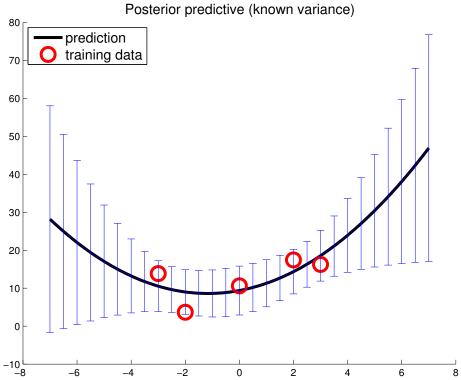
$$\sigma^2_{\text{new}} = \underbrace{x_{\text{new}}^T \Sigma_P x_{\text{new}}}_{\text{from variance in } \theta} + \underbrace{\sigma^2}_{\text{inherent variance in } y} \qquad \text{principled way to handle uncertainty} \tag{58}$$

Prediction uncertainty

Predictive distributions allow mean prediction with error bands.



(a) MLE: constant error bars



(b) Posterior: larger error bars where training points are few

Murphy. Machine Learning: a Probabilistic Perspective, Fig.7.12(a)(b)

Conclusion

Frequentist
- Average over data (both observed and unobserved)
- No principled way to choose estimators
- Less computation

Bayesian
- Average over parameters (subjective prior)
- Uncertainty estimation "for free"
- Computationally intensive

Bayesian methods
1. Specify likelihood / model
2. Choose (conjugate) prior
3. Bayesian inference...