CMPUT 267 Basics of Machine Learning

Winter 2024

UNIVERSITY OF ALBERTA

February 6, 2024

Outline

1. Recap: Parameter Estimation
2. Examples
3. Consistency and Bias
4. Bayesian Approaches

Parameter Estimation

1. Given dataset $\mathcal{D} = \{x_i\}_{i=1}^n$
2. Pick a distribution class (function class, hypothesis space) to model the distribution of x
   - E.g. if $x_i \in \mathbb{R}$, maybe Guassian, $p(x \mid \mathbf{w})$ where $\mathbf{w} = (\mu, \sigma) \in \mathbb{R}^2$

$$p(x \mid \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right).$$

   - E.g. If $x_i \in \{0, 1\}$, Bernoulli $w \in [0, 1]$ where $p(x = 1 \mid w) = w$,

$$p(x \mid w) = w^x (1-w)^{1-x}.$$

3. Identify *best* parameter $\mathbf{w}$ - MLE or MAP estimate

MAP Example, Poisson data with Gamma prior

Suppose we have a dataset $\mathcal{D} = \{8, 4, 5, 9, 5, 2\}$, with each value drawn i.i.d from an unknown Poisson distribution with parameter λ_0. We have a Gamma prior over λ:

$$\textbf{prior}\ p(\lambda) = \frac{\lambda^{k-1} e^{-\lambda/\theta}}{\theta^k \Gamma(k)} \quad \text{and } \textbf{likelihood}\ p(\mathcal{D}|\lambda) = \frac{\lambda^{(\sum_{i=1}^n x_i)} e^{-n\lambda}}{\prod_{i=1}^n x_i!}$$

MAP Example, Poisson data with Gamma prior

Suppose we have a dataset $\mathcal{D} = \{8, 4, 5, 9, 5, 2\}$, with each value drawn i.i.d from an unknown Poisson distribution with parameter λ_0. We have a Gamma prior over λ:

prior $p(\lambda) = \frac{\lambda^{k-1} e^{-\lambda/\theta}}{\theta^k \Gamma(k)}$ and **likelihood** $p(\mathcal{D}|\lambda) = \frac{\lambda^{(\sum_{i=1}^n x_i)} e^{-n\lambda}}{\prod_{i=1}^n x_i!}$

$$p(\mathcal{D}) = \int_0^\infty p(\mathcal{D} \mid \lambda) p(\lambda)\, d\lambda$$

$$= \int_0^\infty \frac{\lambda^{s_n} e^{-n\lambda}}{\prod_{i=1}^n x_i!} \cdot \frac{\lambda^{k-1} e^{-\frac{\lambda}{\theta}}}{\theta^k \Gamma(k)}\, d\lambda$$

$$= \frac{\Gamma(k+s_n)}{\theta^k \Gamma(k) \prod_{i=1}^n x_i! \left(n+\frac{1}{\theta}\right)^{(k+s_n)}}$$

Example (cont'd)

Posterior. $p(\lambda \mid \mathcal{D}) = \frac{p(\mathcal{D}|\lambda)p(\lambda)}{p(\mathcal{D})}$

$$p(\lambda \mid \mathcal{D}) = \frac{\lambda^{s_n} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \cdot \frac{\lambda^{k-1} e^{-\frac{\lambda}{\theta}}}{\theta^k \Gamma(k)} \cdot \frac{\theta^k \Gamma(k) \prod_{i=1}^{n} x_i! \left(n + \frac{1}{\theta}\right)^{(k+s_n)}}{\Gamma\left(k + s_n\right)}$$

Example (cont'd)

Posterior. $p(\lambda \mid \mathcal{D}) = \frac{p(\mathcal{D}|\lambda)p(\lambda)}{p(\mathcal{D})}$

$$p(\lambda \mid \mathcal{D}) = \frac{\lambda^{s_n} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \cdot \frac{\lambda^{k-1} e^{-\frac{\lambda}{\theta}}}{\theta^k \Gamma(k)} \cdot \frac{\theta^k \Gamma(k) \prod_{i=1}^{n} x_i! \left(n + \frac{1}{\theta}\right)^{(k+s_n)}}{\Gamma\left(k + s_n\right)}$$

$$= \frac{\lambda^{((k+s_n)-1)} \cdot e^{-\lambda(n+1/\theta)} \cdot \left(n + \frac{1}{\theta}\right)^{(k+s_n)}}{\Gamma\left(k + s_n\right)}$$

Example (cont'd)

Posterior. $p(\lambda \mid \mathcal{D}) = \frac{p(\mathcal{D}|\lambda)p(\lambda)}{p(\mathcal{D})}$

$$p(\lambda \mid \mathcal{D}) = \frac{\lambda^{s_n} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \cdot \frac{\lambda^{k-1} e^{-\frac{\lambda}{\theta}}}{\theta^k \Gamma(k)} \cdot \frac{\theta^k \Gamma(k) \prod_{i=1}^{n} x_i! \left(n + \frac{1}{\theta}\right)^{(k+s_n)}}{\Gamma\left(k + s_n\right)}$$

$$= \frac{\lambda^{((k+s_n)-1)} \cdot e^{-\lambda(n+1/\theta)} \cdot \left(n + \frac{1}{\theta}\right)^{(k+s_n)}}{\Gamma\left(k + s_n\right)}$$

$$= \frac{\lambda^{((k+s_n)-1)} \cdot e^{-\lambda(n+1/\theta)}}{\left(\frac{1}{n+\frac{1}{\theta}}\right)^{(k+s_n)} \cdot \Gamma\left(k + s_n\right)}$$

Example (cont'd)

Posterior. $p(\lambda \mid \mathcal{D}) = \frac{p(\mathcal{D}|\lambda)p(\lambda)}{p(\mathcal{D})}$

$$p(\lambda \mid \mathcal{D}) = \frac{\lambda^{s_n} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \cdot \frac{\lambda^{k-1} e^{-\frac{\lambda}{\theta}}}{\theta^k \Gamma(k)} \cdot \frac{\theta^k \Gamma(k) \prod_{i=1}^{n} x_i! \left(n + \frac{1}{\theta}\right)^{(k+s_n)}}{\Gamma(k+s_n)}$$

$$= \frac{\lambda^{((k+s_n)-1)} \cdot e^{-\lambda(n+1/\theta)} \cdot \left(n + \frac{1}{\theta}\right)^{(k+s_n)}}{\Gamma(k+s_n)}$$

$$= \frac{\lambda^{((k+s_n)-1)} \cdot e^{-\lambda(n+1/\theta)}}{\left(\frac{1}{n+\frac{1}{\theta}}\right)^{(k+s_n)} \cdot \Gamma(k+s_n)}$$

That is, a Gamma(k', θ') distribution with

$$k' = k + s_n \text{ and } \theta' = \frac{\theta}{n\theta + 1} = \frac{1}{n + 1/\theta}$$

Conjugate Priors

- A conjugate prior $p(w)$ for the parameter of a data distribution $p(x|w)$, where the posterior $p(w|\mathcal{D})$ is of the same type as $p(w)$.
- **Gamma** is a conjugate prior for the parameter of a **Poisson** data distribution
  - Starting from **prior** Gamma(k, θ) and assuming a Poisson **likelihood**, after seeing data $\mathcal{D} = x_1, \ldots, x_n$, the **posterior** is Gamma$\left(k + \sum_{i=1}^{n} x_i, \frac{1}{n+1/\theta}\right)$.
- Similarly, **Beta** is a conjugate prior for the parameter of a **Binomial** data distribution
  - Starting from **prior** Beta(a, b) and assuming a Binomial **likelihood**, after seeing data $\mathcal{D} = n_1$ successes and n_0 failures, the **posterior** is Beta$(a + n_1, b + n_0)$.

Updated data

- What if we observe more data?
- Binomial data example
- We have estimates $a' = a + n_1, b' + n_0$ for posterior Beta(a', b') from data $\mathcal{D}$.
- Now we have additional data $\mathcal{D} \cup x_{n+1}, \ldots, x_{n+10}$.
- Compute $s_{n+10} = s_n + \sum_{i=n+1}^{n+10} x_i$.
- Then $\tilde{a} = a' + \tilde{n}_1$ and $\tilde{b} = b' + \tilde{n}_0$.
- Beta(a', b') is like a new prior

Parameter Estimation: Consistency and Bias

- Estimation of Poisson parameter (estimate λ^*: $x_i \sim p(x|\lambda^*)$):

$$w_{\text{MLE}} = \frac{s_n}{n}, \quad w_{\text{MAP}} = \frac{(k-1) + s_n}{n + 1/\theta}, \qquad s_n = \sum_{i=1}^{n} x_i$$

Point estimates

- MAP and MLE estimates are **point estimates**.
- Suppose we have a dataset $\mathcal{D}$ that was generated by a model:

$$f(\cdot \mid \theta^*) \in \mathcal{F} = \{f(\cdot \mid \theta) \mid \theta \in \mathbb{R}\}$$

- A point estimate answers the question: What is the single best guess for the parameter?
- MLE: $\arg\max_\theta p(\mathcal{D} \mid \theta)$. (mode of the likelihood function)
- MAP: $\arg\max_\theta p(\theta \mid \mathcal{D})$. (mode of the posterior distribution)
- Estimate of θ that has the lowest expected error?

Bayes Estimates

- Bayes estimates estimate the entire posterior distribution, $p(\theta|\mathcal{D})$.
- The posterior is then used in two ways:
  1. Assess the range of plausible parameters given our data, $p(\theta \in [\mu - \epsilon, \mu + \epsilon])$ where μ is the mean of $p(\theta|\mathcal{D})$.
     - $[\mu - \epsilon, \mu + \epsilon]$ is the **credible interval**.
  2. Define an alternate objective for selecting a point estimate: minimize the posterior risk

$$c(\hat{\theta}) = \int_{\mathcal{F}} \ell(\theta, \hat{\theta}) p(\theta \mid \mathcal{D}) \, d\theta,$$

     where $\ell(\theta, \hat{\theta})$ is the *loss*

Bayes Estimator

The Bayes estimator is the point estimate that minimizes the **posterior risk** $c(\hat{\theta})$, where

$$c(\hat{\theta}) = \int_{\mathcal{F}} \ell(\theta, \hat{\theta}) p(\theta \mid \mathcal{D}) \, d\theta$$

The **loss** $\ell(\theta, \hat{\theta})$ expresses how *wrong* we are if we estimate $\hat{\theta}$ when the true answer is θ.

Bayes Estimator for Squared Loss

When $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$:

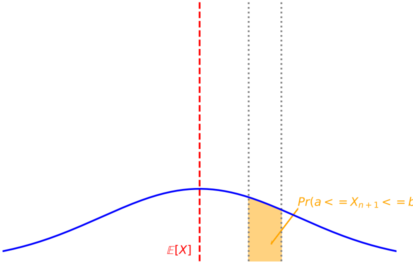
$$c(\hat{\theta}) = \int_{\mathcal{F}} (\theta - \hat{\theta})^2 p(\theta \mid \mathcal{D}) \, d\theta$$

$$\theta_{\text{B}} = \hat{\theta} = \int_{\mathcal{F}} \theta p(\theta \mid \mathcal{D}) \, d\theta = \mathbb{E}[\theta \mid \mathcal{D}]$$

Bayesian Reasoning

- How do we assess our prediction X_{n+1}?
- How do we answer $\Pr(a \leq X_{n+1} \leq b)$?

1. MLE: $F(b \mid \theta_{\text{MLE}}) - F(a \mid \theta_{\text{MLE}})$
2. MAP: $F(b \mid \theta_{\text{MAP}}) - F(a \mid \theta_{\text{MAP}})$
3. Bayes optimal estimator:
   $F(b \mid \theta_{\text{B}}) - F(a \mid \theta_{\text{B}})$
4. Bayesian: $\int_{\mathcal{F}} [F(b \mid \theta) - F(a \mid \theta)]\, p(\theta \mid \mathcal{D})\, d\theta$
   $= \mathbb{E}\left[F(b \mid \theta) - F(a \mid \theta) \,\middle|\, \mathcal{D}\right]$



How do we get model evidence?

To compute the Bayes estimator, we will need the full **posterior** $p(w|\mathcal{D})$ (see slide 12).

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

$$p(\mathcal{D}) = \int p(\mathcal{D}, w)dw = \int p(\mathcal{D}|w)p(w)dw = \mathbb{E}[p(\mathcal{D}|w)]$$

So we need to compute the model evidence $p(\mathcal{D})$ as well. How do we compute $p(\mathcal{D})$?

1. Numerical integration
   $w_1, w_2, \ldots, w_m \sim p(w)$, then $p(\mathcal{D}) = \frac{1}{m}\sum_{i=1}^{m} p(\mathcal{D}|w_i)$
   - as m increases, this approximation gets better.

2. In some cases, we may have a closed-form for the integral. (with the concept of conjugate priors)

Estimation

1. True data distirubtion is p_{true}.

2. Get dataset $\mathcal{D} = \{x_i\}_{i=1}^n$ where X_i has distribution p_{true}.

3. Estimate properties of p_{true}.
   - $\mathbb{E}[X_i]$ or $\text{Var}(X_i)$
   - p_{true} itself.

4. Pick a distribution class to model p_{true}.
   - Gaussian $\mathcal{N}(\mu, \sigma^2 = 1)$, parameter $w = \mu$ to estimate.
   - Poisson with $w = \lambda$.
   - Complex distributions like a misture $p(x) = c_1\mathcal{N}(\mu_1, \sigma_1^2) + c_2\mathcal{N}(\mu_2, \sigma_2^2)$, with $\mathbf{w} = (c_1, \mu_1, \sigma_1^2, c_2, \mu_2, \sigma_2^2)$.

5. Define objective to get w
   - MLE $c(w) = \ln p(\mathcal{D}|w)$
   - MAP $c(w) = \ln p(w|\mathcal{D})$
   - Bayesian: $p(w|\mathcal{D})$

Conditional Models

- We may want to ask questions like "with what probability is some image an image of a cat".
- How can we approach this with what we have been learning?
- We would like something like:

$$\Pr(Y = \text{cat} | X = \mathbf{x})$$

  where $\mathbf{x}$ are the pixels that describe the image.
- Or you might have $\{(x_i, y_i)\}$ and you might want $\Pr(Y = 10 | X = x)$.
- Our models can be parametrized families of conditional distributions

$$\mathcal{F} = \{f(y \mid x; \theta) \mid \theta \in \mathbb{R}^k\}$$

MLE, MAP, Bayesian Prediction for Conditional Models

- Given a hypothesis space $\mathcal{F} = \{p(\cdot \mid \cdot, \theta) \mid \theta \in \mathbb{R}\}$ and a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ of observations $\boldsymbol{x}_i$ and their corresponding targets $\boldsymbol{y}_i$:
- MLE: $p(y|x) = p(y \mid x, \theta_{\text{MLE}})$ where $\theta_{\text{MLE}} = \arg\max_{\theta} \sum_i \ln p(y_i \mid x_i, \theta)$
- MAP: $p(y|x) = p(y \mid x, \theta_{\text{MAP}})$ where $\theta_{\text{MAP}} = \arg\max_{\theta} \ln p(\theta) + \sum_i \ln p(y_i \mid x_i, \theta)$
- Bayesian: $p(y \mid x) = \int_{\mathcal{F}} p(y \mid x, \theta) p(\theta \mid \mathcal{D})\, d\theta$