Recitation 8

Bayesian Methods

DS-GA 1003 Machine Learning

CDS

March 21, 2023

Agenda

1. Recap: MLE
2. Bayesian Methods
3. Questions

MLE for Conditional Probability Models

- Observed data $\mathcal{D} = \{x_{1\ldots n}, y_{1\ldots n}\}$
- Compute likelihood of the data as a function of parameter(s) θ

$$L_{\mathcal{D}}(\theta) = \prod_{i=1}^{n} p(y_i | x_i; \theta)$$

- Find that value of $\theta \in \Theta$ which maximizes the likelihood $\rightarrow$ MLE
  - MLE is the ERM of NLL loss

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \prod_{i=1}^{n} p(y_i | x_i; \theta)$$

- And we make predictions on new points x' as:

$$\hat{f}(x') = p(y | x'; \hat{\theta}_{MLE})$$

MLE for Conditional Probability Models

- Observe that $\hat{\theta}_{MLE}$ is very dependent on the observed data
- Can we do better? What if you have an intuition/belief about the parameter θ before observing the data $\mathcal{D}$?

Bayesian Methods

- Ingredients:
  - **Parameter space** Θ.
  - **Prior**: Distribution $p(\theta)$ on Θ.
  - **Action space** $\mathcal{A}$.
  - **Loss function**: $\ell : \mathcal{A} \times \Theta \rightarrow \mathbb{R}$.

Bayesian Methods

- Ingredients:
  - **Parameter space** Θ.
  - **Prior**: Distribution $p(\theta)$ on Θ.
  - **Action space** $\mathcal{A}$.
  - **Loss function**: $\ell : \mathcal{A} \times \Theta \to \mathbb{R}$.
- The prior $p(\theta)$ represents your belief about the parameter without seeing the data

Bayesian Methods

- Ingredients:
  - **Parameter space** Θ.
  - **Prior**: Distribution $p(\theta)$ on Θ.
  - **Action space** $\mathcal{A}$.
  - **Loss function**: $\ell : \mathcal{A} \times \Theta \to \mathbb{R}$.
- The prior $p(\theta)$ represents your belief about the parameter without seeing the data
- And you update this belief based on observing the data $\mathcal{D}$ with Bayes rule to get the posterior
- Posterior $p(\theta|D) \propto p(\mathcal{D}|\theta)p(\theta)$
- From this distribution, we can get point estimates or take actions

Bayesian Decision Theory

- Ingredients:
  - **Parameter space** Θ.
  - **Prior**: Distribution $p(\theta)$ on Θ.
  - **Action space** $\mathcal{A}$.
  - **Loss function**: $\ell : \mathcal{A} \times \Theta \to \mathbb{R}$.
- 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.**

Bayesian Decision Theory

- Ingredients:
  - **Parameter space** Θ.
  - **Prior**: Distribution $p(\theta)$ on Θ.
  - **Action space** $\mathcal{A}$.
  - **Loss function**: $\ell : \mathcal{A} \times \Theta \to \mathbb{R}$.
- 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:

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

MAP Estimator

- How do we predict y from the posterior of θ?
- MAP estimator for θ from the posterior

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta \mid \mathcal{D})$$

- We can predict y by

$$\hat{y} = \arg\max_{y} p(y \mid x; \theta = \hat{\theta}_{MAP})$$

The Posterior Predictive Distribution

- The MAP estimator only depends on the **mode** of the posterior. Is there a way to incorporate all the information from the posterior?
- 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.
- May not have closed form. Numerical integral may be hard to compute.

MAP Estimator vs Posterior Predictive Distribution

- How do we predict by posterior predictive distribution given a new data point?

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

- Different to the MAP estimator:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta \mid \mathcal{D})$$

$$\hat{y} = \arg\max_{y} p(y \mid x; \theta = \hat{\theta}_{MAP})$$

- In general, the predictions from two methods are different.

MAP Estimator Vs MLE

- MLE looks for the value that maximizes likelihood alone

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L_{\mathcal{D}}(\theta) = \arg\max_{\theta} \prod_{i=1}^{n} p(y_i|x_i;\theta)$$

- MAP maximizes the posterior i.e. a combination of prior and likelihood

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta \mid \mathcal{D}) = \arg\max_{\theta} L_{\mathcal{D}}(\theta)p(\theta)$$

Question 1

Question 1. (From DeGroot and Schervish) Let θ denote the proportion of registered voters in a large city who are in favor of a certain proposition. Suppose that the value of θ is unknown, and two statisticians A and B assign to θ the following different (beta) prior PDFs $\xi_A(\theta)$ and $\xi_B(\theta)$, respectively:

$$\begin{aligned} \xi_A(\theta) &= 2\theta && \text{for } 0 < \theta < 1, \\ \xi_B(\theta) &= 4\theta^3 && \text{for } 0 < \theta < 1. \end{aligned}$$

In a random sample of 1000 registered voters from the city, it is found that 710 are in favor of the proposition.

- Find the posterior distribution that each statistician assigns to θ.

Question 1: Solution

- Likelihood of the observed data, 710 in-favour, 290 against:

$$f(x|\theta) = \theta^{710}(1-\theta)^{290}$$

- Multiplying by the two priors ξ_A and ξ_B, we have

$$\xi_A(\theta|x) \propto f(x|\theta)\xi_A(\theta) \propto \theta^{711}(1-\theta)^{290}$$

and

$$\xi_B(\theta|x) \propto f(x|\theta)\xi_B(\theta) \propto \theta^{713}(1-\theta)^{290}.$$

Question 1: Solution

- Multiplying by the two priors ξ_A and ξ_B, we have

$$\xi_A(\theta|x) \propto f(x|\theta)\xi_A(\theta) \propto \theta^{711}(1-\theta)^{290}$$

and

$$\xi_B(\theta|x) \propto f(x|\theta)\xi_B(\theta) \propto \theta^{713}(1-\theta)^{290}.$$

- Thus the posteriors from A and B are both beta with parameters $(712, 291)$ and $(714, 291)$, respectively.

Question 1

Question 1. (From DeGroot and Schervish) Let θ denote the proportion of registered voters in a large city who are in favor of a certain proposition. Suppose that the value of θ is unknown, and two statisticians A and B assign to θ the following different prior PDFs $\xi_A(\theta)$ and $\xi_B(\theta)$, respectively:

$$\begin{aligned} \xi_A(\theta) &= 2\theta \quad \text{for } 0 < \theta < 1, \\ \xi_B(\theta) &= 4\theta^3 \quad \text{for } 0 < \theta < 1. \end{aligned}$$

In a random sample of 1000 registered voters from the city, it is found that 710 are in favor of the proposition.

- Find the Bayes estimate of θ (minimizer of posterior expected loss) for each statistician based on the squared error loss function.

Question 1: Solution

If the loss function is square loss, the minimizer $f^* = E[Y|X]$.

- We have found the two posteriors $\xi_A(\theta|x)$ and $\xi_B(\theta|x)$
- The posteriors from A and B are both beta with parameters $(712, 291)$ and $(714, 291)$, respectively.

Question 1: Solution

If the loss function is square loss, the minimizer $f^* = E[Y|X]$.

- We have found the two posteriors $\xi_A(\theta|x)$ and $\xi_B(\theta|x)$
- The posteriors from A and B are both beta with parameters $(712, 291)$ and $(714, 291)$, respectively.
- Thus minimizers of the posterior expected loss is the respective means are $\frac{712}{1003}$ and $\frac{714}{1005}$.
  - Recall the mean of a Beta distribution $\mathbb{E}[x; a, b] = \frac{a}{a+b}$

Question 2

What would be the Maximum a Posteriori (MAP) estimator for λ for i.i.d. $\{x_1, x_2, \ldots, x_N\}$ where $x_i \sim \exp(\lambda)$ with prior $\lambda \sim \text{Uniform}[u_0, u_1]$?

Question 2: Solution

Question 2: Solution

- Likelihood: $L(x_1, \ldots, x_N|\lambda) = \lambda^N e^{-\lambda(x_1+\cdots+x_N)}$
- log-likelihood: $\ell(\lambda|x_1, \ldots, x_N) = N \ln \lambda - \lambda(x_1 + \cdots + x_N)$
- $\ell'(\lambda) =$

$$\frac{N}{\lambda} - (x_1 + \cdots + x_N) \begin{cases} > 0 & \text{if } 0 < \lambda < 1/\bar{x} = N/(x_1 + \cdots + x_N), \\ = 0 & \text{if } \lambda = 1/\bar{x} \\ < 0 & \text{if } \lambda > 1/\bar{x} \end{cases}$$

- Prior: $p(\lambda) = \frac{1}{u_1 - u_0} \mathbb{1}_{[u_0, u_1]}(\lambda)$.
- Posterior:
  $p(\lambda|x_1, \ldots, x_N) \propto L(x_1, \ldots, x_N|\lambda)p(\lambda) = \lambda e^{-\lambda(x_1+\cdots+x_N)} \mathbb{1}_{[u_0, u_1]}(\lambda)$
- Maximum value of posterior is attained at

$$\lambda = \begin{cases} u_0 & \text{if } u_0 > 1/\bar{x}, \\ 1/\bar{x} & \text{if } u_0 \leq 1/\bar{x} \leq u_1 \\ u_1 & \text{if } u_1 < 1/\bar{x}. \end{cases}$$

Takeaways

- In Bayesian methods, we have a prior that encodes our belief without the data
- We update the prior based on the observed data i.e. likelihood and get the posterior distribution
- What can we do with this distribution? MAP estimator, Bayesian point estimation, credible set, etc.