Bayesian inference

STATS 303 Statistical Machine Learning

Spring 2022

Lecture 4

score function

(log-)

- Given a likelihood function $\mathcal{L}(\theta|\mathcal{X}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta)$, we define the **(Fisher) score function** to be

$$\mathcal{S}(\theta|\mathcal{X}) := \frac{\partial \mathcal{L}(\theta|\mathcal{X})}{\partial \theta}$$

- For MLE, $\mathcal{S}(\hat{\theta}_{\mathrm{MLE}}|\mathcal{X}) = 0.$

- Fact: $\mathbb{E}[\mathcal{S}(\theta|\mathcal{X})] = 0$ (why?).

Fisher information

- The **Fisher information** is defined to be the variance of the score function:

$$\mathcal{I}(\theta) := \mathrm{Var}\big(\mathcal{S}(\theta|\mathcal{X})\big) \overset{\text{(why?)}}{=} -\mathbb{E}\left[\frac{\partial^2 \mathcal{L}(\theta|\mathcal{X})}{\partial\theta^2}\right]$$

Fisher information

- **Remark.** It makes sense to talk about the score function and the Fisher information of a single observation $\boldsymbol{x}_n$ in the sample. We just need to replace $\mathcal{X}$ by $\boldsymbol{x}_n$ in the definitions.

- That is,

$$\mathcal{S}(\theta|\boldsymbol{x}_n) := \frac{\partial \overset{\log}{p}(\boldsymbol{x}_n|\theta)}{\partial \theta}$$

and

$$\mathcal{I}(\theta|\boldsymbol{x}_n) := \mathrm{Var}\big(\mathcal{S}(\theta|\boldsymbol{x}_n)\big) = -\,\mathbb{E}\left[\frac{\partial^2 \overset{\log}{p}(\boldsymbol{x}_n|\theta)}{\partial \theta^2}\right]$$

the Bayesian estimator

previously, ML estimator of the density

$$\hat{p}_{\mathrm{MLE}}(\boldsymbol{x}) = p(\boldsymbol{x}|\hat{\theta}_{\mathrm{MLE}})$$

where $\hat{\theta}_{\mathrm{MLE}}$ is solved by maximizing the (log-)likelihood.

the Bayesian view

- Before looking at a sample, we may have some prior knowledge on the parameter θ. That is, we have a prior density $p(\theta)$
- If the parameter of our model, θ, is **regarded as a random variable**, then we can determine
  - the prior $p(\theta)$
  - the likelihood $p(\mathcal{X}|\theta)$
  - the posterior $p(\theta|\mathcal{X})$
  - the joint probability $p(\mathcal{X},\theta) = p(\mathcal{X}|\theta)p(\theta)$

Bayes' density estimation

- Combining $p(\theta)$ with the likelihood density $p(\mathcal{X}|\theta)$, we have, by Bayes' rule,

$$p(\theta|\mathcal{X}) = \frac{p(\mathcal{X}|\theta)p(\theta)}{p(\mathcal{X})} = \frac{p(\mathcal{X}|\theta)p(\theta)}{\int p(\mathcal{X}|\theta')p(\theta')d\theta'}$$

- We have

$$\hat{p}_{\mathrm{Bayes}}(\boldsymbol{x}) = p(\boldsymbol{x}|\mathcal{X}) = \int p(\boldsymbol{x},\theta|\mathcal{X})d\theta$$

$$= \int p(\boldsymbol{x}|\theta,\mathcal{X})p(\theta|\mathcal{X})d\theta = \int p(\boldsymbol{x}|\theta)p(\theta|\mathcal{X})d\theta$$

If only I had a way to integrate …

- $\hat{p}_{\text{Bayes}}(\boldsymbol{x}) = \int p(\boldsymbol{x}|\theta) \frac{p(\mathcal{X}|\theta)p(\theta)}{\int p(\mathcal{X}|\theta')p(\theta')d\theta'} d\theta$
- Numerical integration methods do not work if we are in high dimension.
- We will need to use sampling methods, which we will discuss later in the course.

Bayes' density estimation

Bayesian inference

subjective

$$\hat{p}_{\text{Bayes}}(\boldsymbol{x}) = \int p(\boldsymbol{x}|\theta)p(\theta|\mathcal{X})d\theta$$

frequentist inference

objective

$$\hat{p}_{\text{MLE}}(\boldsymbol{x}) = p(\boldsymbol{x}|\hat{\theta}_{\text{MLE}})$$

Bayes' density estimation

- We have

$$\hat{p}_{\text{Bayes}}(\boldsymbol{x}) = \int p(\boldsymbol{x}|\theta)p(\theta|\mathcal{X})d\theta,$$

which is usually difficult to evaluate.

- Suppose $p(\theta|\mathcal{X})$ is concentrated around a single point $\hat{\theta}$. Then

$$\int p(x|\theta)p(\theta|\mathcal{X})d\theta = \mathbb{E}_{\theta\sim p(\theta|\mathcal{X})}[p(x|\theta)] \approx p(x|\hat{\theta}).$$

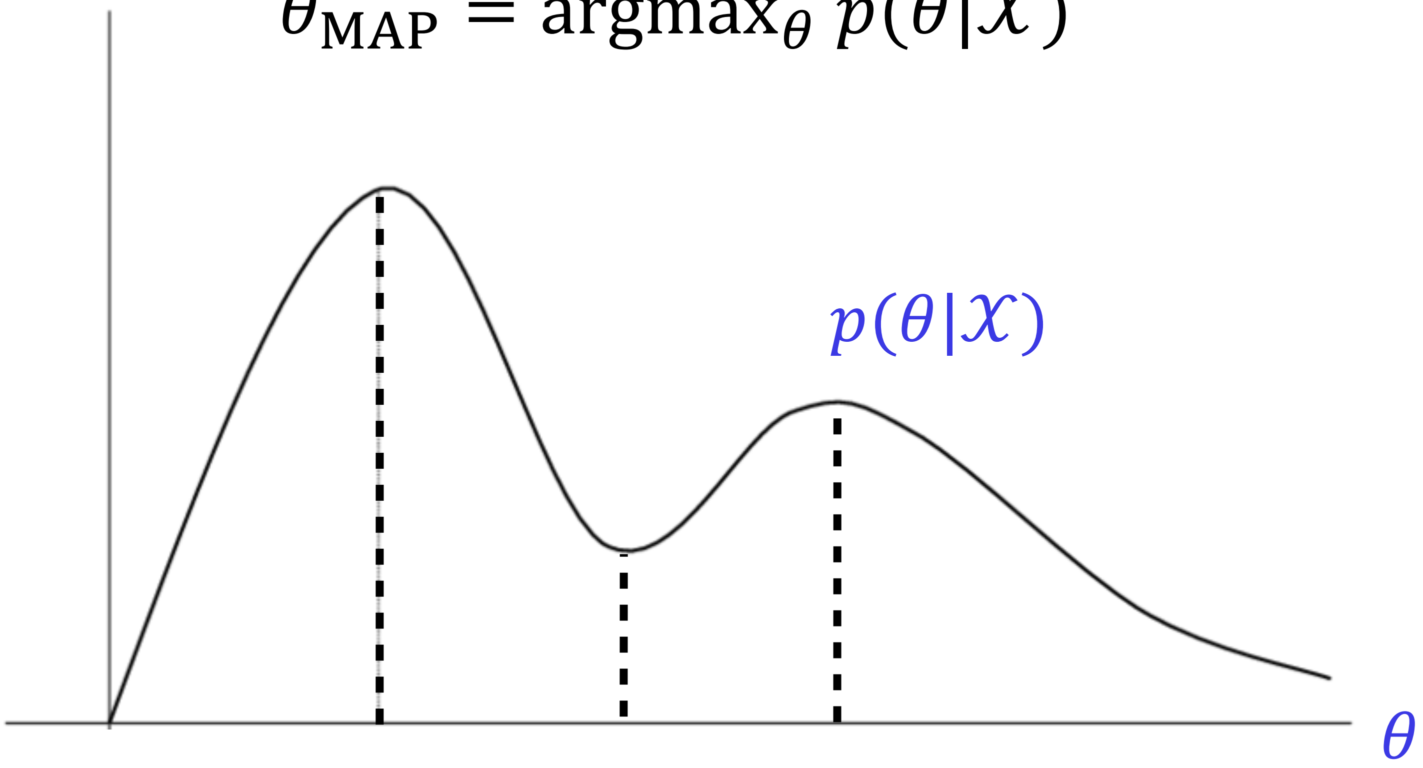
- We recover the ML density estimator if, in the above, $\hat{\theta} = \hat{\theta}_{\text{MLE}}$. However, that is far from the idea of concentration of $p(\theta|\mathcal{X})$.

Bayes' density estimation

MAP: maximum a posteriori

- Another choice: let $\hat{\theta} = \hat{\theta}_{\mathrm{MAP}}$ be determined by

$$\hat{\theta}_{\mathrm{MAP}} = \mathrm{argmax}_{\theta}\, p(\theta|\mathcal{X})$$



Bayes' density estimation

- Let's look at a Gaussian example of MAP.

Consider $\mathcal{N}(x \mid \theta, \sigma^2)$ where σ^2 is known and fixed.

Also $p(\theta) = \mathcal{N}(\theta \mid 0, \sigma_0^2) = \frac{1}{\sqrt{2\pi}\,\sigma_0} \exp\left(-\frac{\theta^2}{2\sigma_0^2}\right)$

Given a sample $\mathcal{X} = \{x_n\}_{n=1}^{N}$ i.i.d. from $\mathcal{N}(x \mid \theta, \sigma^2)$.

The likelihood

$$
\begin{aligned}
p(\mathcal{X} \mid \theta) &= \prod_{n=1}^{N} \mathcal{N}(x_n \mid \theta, \sigma^2) \\
&= \prod_{n=1}^{N} \left(\frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x_n-\theta)^2}{2\sigma^2}\right)\right) \\
&= \frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\theta)^2\right)
\end{aligned}
$$

The posterior

$$
p(\theta \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \theta)\, p(\theta)}{\int p(\mathcal{X} \mid \theta')\, p(\theta')\, d\theta'} \propto p(\mathcal{X} \mid \theta)\, p(\theta)
$$

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

$$
= \arg\max_{\theta} p(\mathcal{X} \mid \theta)\, p(\theta) = \arg\max_{\theta} \log p(\mathcal{X} \mid \theta) + \log p(\theta)
$$

$$
= \arg\max_{\theta} \underbrace{\left[-\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\theta)^2 - \frac{1}{2}\log(2\pi) - \log\sigma_0 - \frac{\theta^2}{2\sigma_0^2} \right]}_{\text{"}f(\theta)\text{"}}
$$

Setting

$$\frac{df(\theta)}{d\theta} = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \theta) - \frac{\theta}{\sigma_0^2} = 0,$$

we have

$$\sigma_0^2\left(\sum_{n=1}^{N} x_n - N\theta\right) = \sigma^2\theta.$$

That is,

$$(N\sigma_0^2 + \sigma^2)\,\theta = \sigma_0^2 \sum_{n=1}^{N} x_n$$

Hence,

$$\theta = \frac{\sum_{n=1}^{N} x_n}{N + \frac{\sigma^2}{\sigma_0^2}}$$

Conclusion:

$$\hat{\theta}_{MAP} = \frac{\sum_{n=1}^{N} x_n}{N + \frac{\sigma^2}{\sigma_0^2}}$$

Bayes' density estimation

- Yet another choice other than MAP:

$$\hat{\theta}_{\mathrm{Bayes}} = \mathbb{E}[\theta|\mathcal{X}] = \int \theta p(\theta|\mathcal{X}) d\theta$$

How do we choose the prior $p(\theta)$?

- If $p(\theta)$ is chosen so that $p(\theta|\mathcal{X})$ is of the same parametric form as $p(\theta)$, then $p(\theta)$ is said to be a **conjugate prior** for the likelihood $p(\mathcal{X}|\theta)$.
- For a Gaussian likelihood, we can choose a Gaussian conjugate prior (exercise).
- Let's work with the case of multinomial variable.

conjugate prior for multinomial likelihood

Recall: X takes state i with probability q_i, $i=1,\cdots,K$.

Each observation $x_n \in \mathbb{R}^K$ is a one-hot vector.

- Parameters: $q = (q_1, \cdots, q_K)^T$

- Likelihood: $p(x|q) = \prod_{n=1}^{N} \prod_{i=1}^{K} q_i^{x_{ni}}$

Take the prior to be a Dirichlet distribution

$$p(q) = \mathrm{Dir}(q \mid \alpha) := \boxed{\frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)}} \prod_{i=1}^{K} q_i^{\alpha_i - 1}$$

where $\alpha = (\alpha_1, \cdots, \alpha_K)^T$ ← nothing more than a normalization factor

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} \, du$$

Remark: $\Gamma(n) = (n-1)!$, $n = 1, 2, 3, \cdots$

Given the likelihood and the prior, the posterior is

$$p(q \mid X) \propto p(X \mid q) \, p(q)$$

$$= \prod_{n=1}^{N} \prod_{i=1}^{K} q_i^{x_{ni}} \cdot \prod_{i=1}^{K} q_i^{\alpha_i - 1}$$

$$= \prod_{i=1}^{K} q_i^{\sum_{n=1}^{N} x_{ni} + \alpha_i - 1}$$

Let N_i denote the number of times we see state i in X.

Then $N_i = \sum_{n=1}^{N} x_{ni}$.

We have $p(q \mid X) \propto \prod_{i=1}^{K} q_i^{(\alpha_i + N_i) - 1}$,

which is of the same parametric form as $p(q)$.

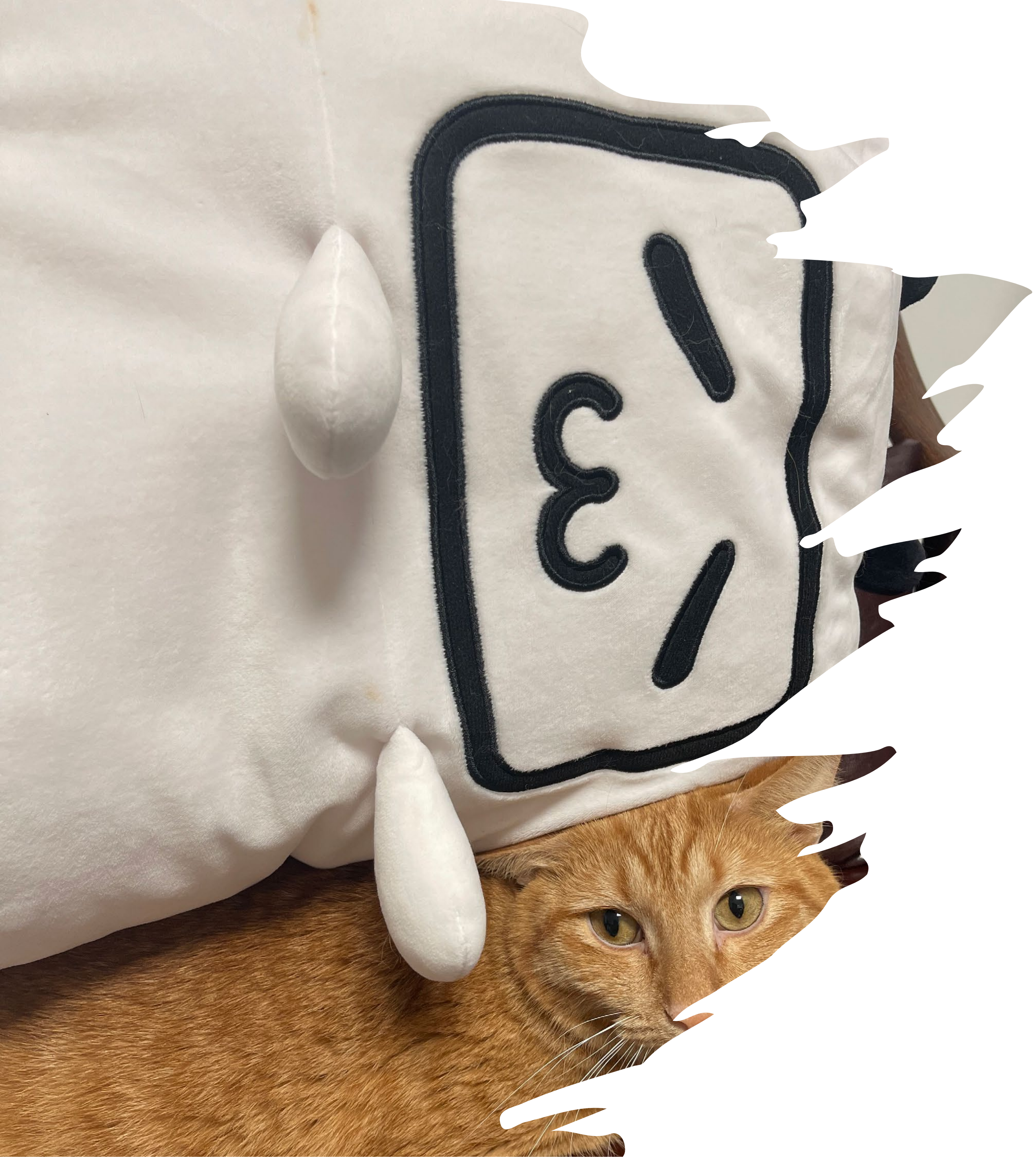
Therefore, we say that $\text{Dir}(q|\alpha)$ is a <u>conjugate prior</u> for the multinomial likelihood $p(X|q)$.

Remark: We can easily get the normalization factor by looking at the prior:

$$p(q|X) = \frac{\Gamma\left(\sum_{i=1}^{K}(\alpha_i+N_i)\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i+N_i)} \prod_{i=1}^{N} q_i^{(\alpha_i+N_i)-1}$$

Intuition:

$$\text{prior}_1 \xrightarrow{\text{data}} \begin{matrix}\text{posterior}_1\\ \text{prior}_2\end{matrix} \xrightarrow{\text{data}} \begin{matrix}\text{posterior}_2\\ \text{prior}_3\end{matrix} \longrightarrow \cdots$$



Questions?

Reference

- *Bayesian inference:*
    - *[Al] Ch.4.4, 16.1, 16.2*
    - *[Bi] Ch.2.2.1 (for Dirichlet distribution)*
    - *[HaTF] Ch.8.3*
- *Parametric classification and regression:*
    - *[Al] Ch.4.5*