Bayesian inference

Corrado Possieri

Machine and Reinforcement Learning in Control Applications

Introduction

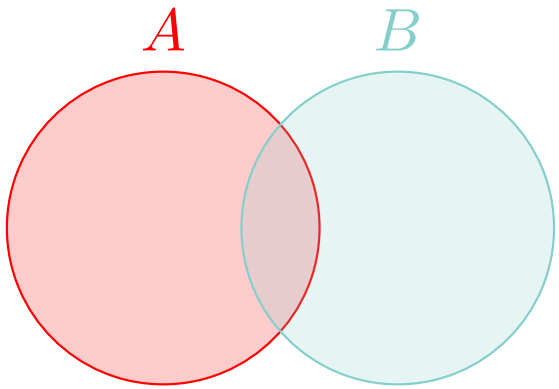
- Bayesian inference provides the tools to update one's beliefs in the evidence of new data.
- The Bayes' theorem states that

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)},$$

where

  - A is some proposition about the world,
  - B is some data or evidence.

Bayes' theorem



- $p(A \cap B)$ is the joint probability of A and B, which satisfies

$$p(A \cap B) = p(A|B)p(B) = p(B|A)p(A),$$

- Thus, we have that

$$p(A|B) = \frac{p(A \cap B)}{p(B)} = \frac{p(B|A)p(A)}{p(B)},$$

that is the Bayes' theorem.

Law of total probability

- Let $\{A_i, i = 1, 2, \dots\}$ be a partition of a sample space.
- Suppose that each event A_i is measurable.
- For any event A of the same probability space, it holds that

$$p(B) = \sum_i p(B|A_i)p(A_i).$$

- Bayes' rule can be rewritten as

$$p(A_i|B) = \frac{p(B|A_i)p(A_i)}{\sum_i p(B|A_i)p(A_i)},$$

A simple example

- What is the probability that it rained given the observation that is wet?

$$p(\text{rain}|\text{wet}) = \frac{p(\text{wet}|\text{rain})p(\text{rain})}{p(\text{wet}),}$$

  - $p(\text{rain}|\text{wet})$: probability that it rained given that it is wet;
  - $p(\text{rain})$: probability of rain, before looking at the ground;
  - $p(\text{wet}|\text{rain})$: likelihood of the sidewalk being wet, under the assumption that it rained;
  - $p(\text{wet})$: total plausibility of the evidence

$$p(\text{wet}) = p(\text{wet}|\text{rain})p(\text{rain}) + p(\text{wet}|\text{no rain})p(\text{no rain}).$$

- Update our initial beliefs with some observation, yielding a final measure of the plausibility, given the evidence.

Bayes' rule for probability density functions

- Let $\theta \in \mathbb{R}^n$ be a set of parameters we want to estimate.
- The a priori knowledge about θ, before observing any data, is described in the form of a *prior* probability distribution $p(\theta)$.
- Let $\mathcal{D}$ denote a set of observed data related to θ.
- The statistical model relating $\mathcal{D}$ to θ is the *likelihood* $p(\mathcal{D}|\theta)$.
- The Bayes' rule states that

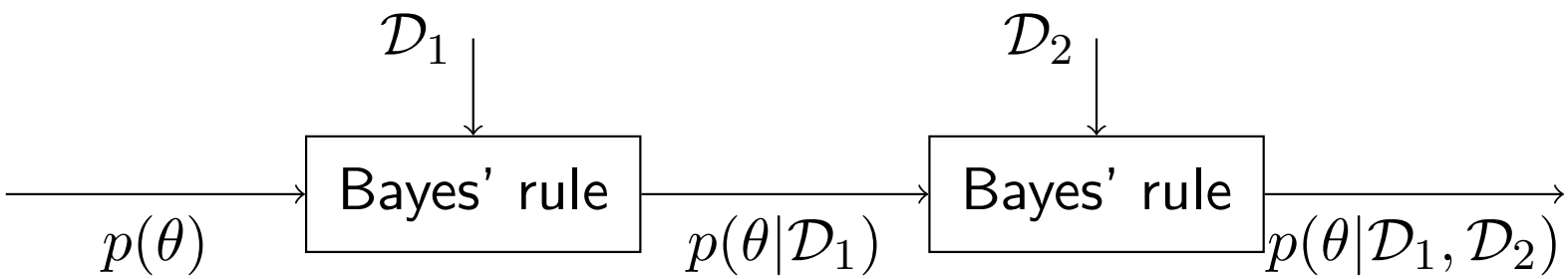
$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)}.$$

- In other words, Bayes' rule can be stated as

$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

Recursive inference

- The Bayesian approach is suitable for on-line inference
  - as new data is gathered the current posterior becomes the new prior, and a new posterior is computed based on the new data.



- The output of this procedure is our belief about the parameters θ given the data $\mathcal{D}_1, \mathcal{D}_2, \ldots$.

Bayesian Estimators

- We want to calculate a single value (known as a statistic) serving as a 'best estimate" of an unknown parameter θ.
- Bayesian point-estimators are the central-tendency statistics of the posterior distribution
  - the posterior mean, which minimizes the (posterior) risk (expected loss) for a squared-error loss function;
  - the posterior median, which minimizes the posterior risk for the absolute-value loss function;
  - the Maximum A Posteriori (MAP) estimate, which finds a maximum of the posterior distribution;
  - the Maximum Likelihood (ML) estimate, which coincides with the MAP under uniform prior probability.

Typical loss functions

- A loss function $\mathcal{L}$ measures the error in predicting θ via $\hat{\theta}$.
- Typical loss functions are
  - quadratic
$$\mathcal{L}(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2;$$
  - linear
$$\mathcal{L}(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|;$$
  - hit-or-miss
$$\mathcal{L}(\theta, \hat{\theta}) = \begin{cases} 1, & \text{if } \|\theta - \hat{\theta}\| > \delta, \\ 0, & \text{if } \|\theta - \hat{\theta}\| \leq \delta; \end{cases}$$
  - Huber loss
$$\mathcal{L}(\theta, \hat{\theta}) = \begin{cases} \delta(\|\theta - \hat{\theta}\| - \frac{\delta}{2}), & \text{if } \|\theta - \hat{\theta}\| > \delta, \\ \frac{1}{2}\|\theta - \hat{\theta}\|^2, & \text{if } \|\theta - \hat{\theta}\| \leq \delta. \end{cases}$$

Estimation

- Given the posterior, Bayesian estimators attempt at determining the value of $\hat{\theta}$ that minimizes the expected loss

$$\begin{aligned}\hat{\theta} &= \arg\min_{\hat{\theta}} \int \mathcal{L}(\theta, \hat{\theta}) p(\theta|\mathcal{D}) \mathrm{d}\theta \\ &= \arg\min_{\hat{\theta}} \mathbb{E}\{\mathcal{L}(\theta, \hat{\theta})|\mathcal{D}\}.\end{aligned}$$

- We focus on the loss function presented in the previous slide.

Minimum Mean-Square Error (MMSE)

- Considering $\mathcal{L}(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$, we want to minimize

$$\int \|\theta - \hat{\theta}\|^2 p(\theta|\mathcal{D}) \mathrm{d}\theta.$$

- The derivative with respect to $\hat{\theta}$ of this error is

$$\frac{\partial}{\partial \hat{\theta}} \int \|\theta - \hat{\theta}\|^2 p(\theta|\mathcal{D}) \mathrm{d}\theta = -\int 2(\theta - \hat{\theta}) p(\theta|\mathcal{D}) \mathrm{d}\theta.$$

- Setting this derivative to zero, we have

$$0 = \int 2(\theta - \hat{\theta}) p(\theta|\mathcal{D}) \mathrm{d}\theta \iff \hat{\theta} = \int \theta p(\theta|\mathcal{D}) \mathrm{d}\theta,$$

- Therefore

$$\hat{\theta}_{\mathrm{MMSE}} = \mathbb{E}\{p(\theta|\mathcal{D})\},$$

i.e., $\hat{\theta}_{\mathrm{MMSE}}$ is the mean of the posterior.

Minimum Absolute Error (MAE)

- In the linear loss $\mathcal{L}(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|$, we want to minimize

$$\int \|\theta - \hat{\theta}\| p(\theta|\mathcal{D}) \mathrm{d}\theta$$

- The minimum of such a function is attained for

$$\int_{-\infty}^{\hat{\theta}} p(\theta|\mathcal{D}) \mathrm{d}\theta = \int_{\hat{\theta}}^{\infty} p(\theta|\mathcal{D}) \mathrm{d}\theta$$

i.e., $\hat{\theta}_{\mathrm{MAE}}$ is the median of the posterior.

Maximum A-Posteriori Estimator (MAP)

- For the hit-or-miss loss, we have

$$\mathbb{E}\{\mathcal{L}(\theta,\hat{\theta})|\mathcal{D}\} = \int_{-\infty}^{\hat{\theta}-\delta} p(\theta|\mathcal{D})\mathrm{d}\theta + \int_{\hat{\theta}+\delta}^{\infty} p(\theta|\mathcal{D})\mathrm{d}\theta$$
$$= 1 - \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} p(\theta|\mathcal{D})\mathrm{d}\theta.$$

- For small δ and suitable assumptions on $p(\theta|\mathcal{D})$, the maximum occurs at the maximum of $p(\theta|\mathcal{D})$, *i.e.*,

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max p(\theta|\mathcal{D}).$$

- Similarly, the maximum likelihood estimator is

$$\hat{\theta}_{\mathrm{ML}} = \arg\max p(\mathcal{D}|\theta),$$

which matches $\hat{\theta}_{\mathrm{MAP}}$ under uniform prior.

Fairness of a coin

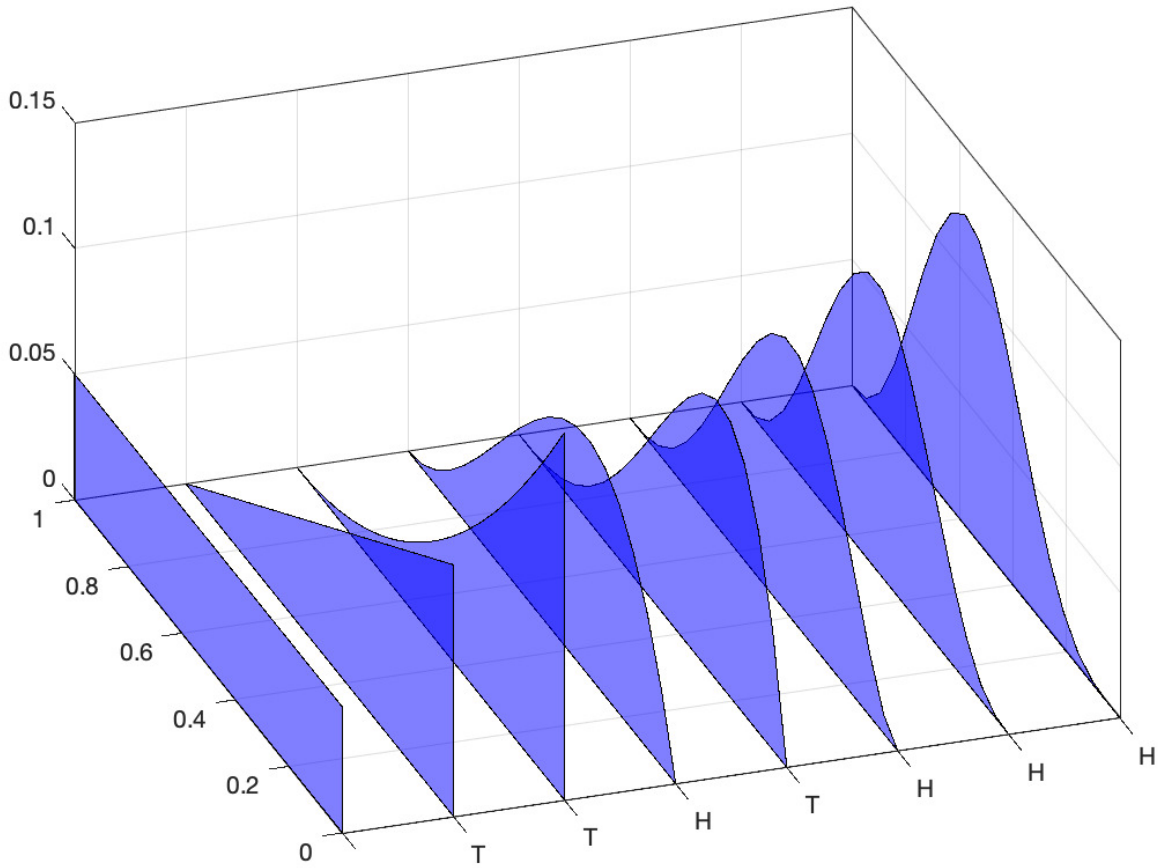
- Suppose we want to estimate the fairness of a coin.
- The evidence is the coin tossing outcome

$$X \in \{\mathsf{H}, \mathsf{T}\}.$$

- We want to estimate the coin bias parameter $\theta \in [0, 1]$.
- The likelihood is

$$p(X = x|\theta) = \begin{cases} \theta, & x = \mathsf{H}, \\ 1 - \theta & x = \mathsf{T}. \end{cases}$$

Fairness updates



Parameter estimation

- Consider the dynamical system

$$\dot{x} = f(x, \theta), \qquad y = h(x).$$

- We want to estimate θ, given the noisy measurements

$$\hat{y}(t_i) = y(t_i) + \delta.$$

- Letting $\phi(t, \theta, x_0)$ be the solution to the system and assuming $\delta \in \mathcal{N}(0, \sigma)$, we have

$$p(\hat{y}|\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\|\hat{y}(t_i) - \phi(t_i, \theta, x_0)\|^2}{\sigma^2}\right).$$

- Using recursively Bayes' formula, we can estimate θ.

Naive Bayes' classifier

- We want classify vectors of n features

$$x = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$$

  into K classes $C_1, \ldots, C_K$.
- The output y is a categorical variable in $\{1, ..., K\}$.
- The classifier assigns to each input vector x the probability

$$p(y|x), \quad y = 1, \ldots, K.$$

- This requires us to specify the conditional distribution $p(x|y)$.
- The simplest approach is to assume the features are conditionally independent given the class label.

$$p(x|y = c, \theta) = \prod_{j=1}^{n} p(x_j|y = c, \theta_{jc}).$$

Naive Bayes' classifier

- In the case of binary features $x_j \in \{0, 1\}$, we can use

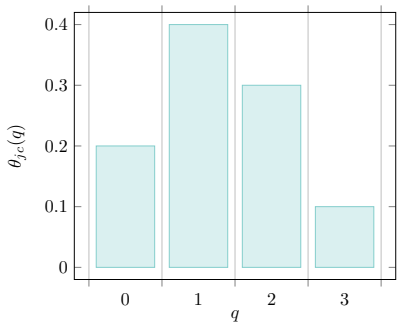
$$p(x_j|y = c, \theta_{jc}) = \text{Ber}(x_j, \theta_{jc}),$$

where θ_{jc} is the probability that feature j occurs in class c.

- In the case of categorical features $x_j \in \{1, ..., Q\}$, we can use

$$p(x_j|y = c, \theta_{jc}) = \prod_{j=1}^{n} \text{Cat}(x_j, \theta_{jc}),$$

that is, if $x_j \sim \text{Cat}(x_j, \theta_{jc})$, then $p(x_j = q|\theta_{jc}) = \theta_{jc}(q)$.



Naive Bayes' classifier

- The class probability, given the input, can be expressed as

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)} = \frac{\prod_{j=1}^{n} p(x_j|C_k)\, p(C_k)}{\sum_{k=1}^{K} \prod_{j=1}^{n} p(x_j|C_k)\, p(C_k)}.$$

- In the case of binary features $x_j \in \{0, 1\}$, we obtain the Bernoulli naive Bayes classifier

$$p(x_j|C_k) = \theta_{jc}^{x_j}(1 - \theta_{jc})^{1-x_j},$$

where θ_{jc} is the probability of class C_k generating x_i, *i.e.*,

$$\theta_{jc} = p(x_j = 1|y = C_k).$$

- The decoupling of the class conditional distributions means that each distribution can be independently estimated.

Training the classifier

- The marginal class probabilities $p(C_1), \ldots, p(C_K)$ are evaluated as the empirical frequencies of the classes

$$p(C_i) = \frac{\text{number of times } y = C_k}{\text{total number of data}}.$$

- The class-conditional probability $p(x_i|C_k)$ is evaluated as the empirical frequency of the outcome x_i for the class C_k

$$p(x_i|C_k) = \frac{\text{number of times } x_i = 1 \text{ in data with } y = C_K}{\text{total number of data with } y = C_k}.$$

The case of two classes

- Suppose that there are just two classes C_1 and C_2.
- Then we have that

$$p(C_1|x) = \frac{p(C_1)}{p(x)} \prod_{i=1}^{n} p(x_i|C_1),$$

$$p(C_2|x) = \frac{p(C_2)}{p(x)} \prod_{i=1}^{n} p(x_i|C_2).$$

- By dividing the two probabilities, we obtain the odds

$$\frac{p(C_1|x)}{p(C_2|x)} = \frac{p(C_1)}{p(C_2)} \frac{\prod_{i=1}^{n} p(x_i|C_1)}{\prod_{i=1}^{n} p(x_i|C_2)} = \frac{p(C_1)}{p(C_2)} \prod_{i=1}^{n} \frac{p(x_i|C_1)}{p(x_i|C_2)}.$$

- D is classified as C_1 if the odds are $\geqslant 1$, or 0 otherwise.