Introduction to Bayesian learning

Juho Lee

The graduate school of AI, KAIST

Learning

- We are given a set of observed data assumed to be generated from some distribution.

$$X := (x_1, \ldots, x_n) \overset{\text{i.i.d.}}{\sim} p_{\text{true}}(x). \tag{1}$$

- Since we don't have an access to $p_{\text{true}}(x)$, we setup a model $p(x; \theta)$ defined with a parameter θ.
- Now we select θ that best describes the observed data X through $p(x; \theta)$.

Learning

- Best describes? - $p(x;\theta)$ should be close to $p_{\text{true}}(x)$.
- A popular example - maximum likelihood.

$$
\begin{aligned}
\mathbb{D}_{\text{KL}}[p_{\text{true}}(x) \| p(x;\theta)] &= \int p_{\text{true}}(x) \log \frac{p_{\text{true}}(x)}{p(x;\theta)} \mathrm{d}x \\
&= -\mathbb{E}_{p_{\text{true}}(x)}[\log p(x;\theta)] - \mathbb{H}[p_{\text{true}}(x)] \\
&\approx -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i;\theta) + \text{const.} \qquad (2)
\end{aligned}
$$

Model

- A simplified representation of (random) phenomenon with mathematical language.
- "All models are wrong, but some are useful." - George E. P. Box.
- How do we know whether a model is good enough?
- How can we compare different models?

Bayesian learning

- It all began from a simple formula.

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}. \tag{3}$$

- Bayesian learning: treat θ as a random variable with prior $p(\theta)$, and compute its posterior $p(\theta|X)$ after observing data.

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} = \frac{p(\theta)\prod_{i=1}^{n} p(x_i|\theta)}{p(X)}. \tag{4}$$

Frequentism vs Bayesianism

- Frequentists
  - Probability is a limiting frequency of an event happening over repeated experiments.
  - Parameter θ is a fixed value, and it is meaningless to define the frequency of θ (and thus $p(\theta)$).
  - We are interested in doing repeated experiments for X (even if it is hypothetical).
- Bayesian
  - Probability is quantification of uncertainty for some event.
  - It is natural to define an uncertainty of a parameter θ as $p(\theta)$.
  - We are interested in the uncertainty of θ after observing data X - the posterior $p(\theta|X)$.
  - We are not particularly interested in the uncertainty of X because we have observed it.

Coin toss example

- Say we have observed a set of outcomes from a coin toss.

$$X = (x_1, \dots, x_n), \quad x_i \in \{\mathsf{H}, \mathsf{T}\} \text{ for } i = 1, \dots, n. \tag{5}$$

- We assume a very simple Bernoulli model.

$$p(x = \mathsf{H}; \theta) = \theta, \quad \theta \in [0, 1]. \tag{6}$$

- We want to estimate the parameter θ.

Coin toss example - a frequentist approach

- Believing that our simple model is correct, there should be only one parameter θ that could have generated X.
- Define an estimator $\hat{\theta}_X$ by maximizing the log-likelihood.

$$\hat{\theta}_X := \arg\max_{\theta} \sum_{i=1}^{n} \log p(x_i; \theta) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{x_i = \mathsf{H}\}}}{n}. \tag{7}$$

- $\hat{\theta}_X$ would approach θ as we observe more and more data (do more coin tosses).

Coin toss example - a frequentist approach

- It is perfectly natural to define a probability of $\hat{\theta}_X$, because we can do repeated experiments to compute them.
- In other words, $\hat{\theta}_X$ itself is a random variable, with mean and variance computed as

$$\mathbb{E}_{p_{\mathrm{true}}(X)}[\hat{\theta}_X] = \theta, \quad \mathrm{Var}_{p_{\mathrm{true}}(X)}[\hat{\theta}_X] = \frac{\theta(1-\theta)}{n}. \tag{8}$$

- By the central limit theorem,

$$\frac{\hat{\theta}_X - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1). \tag{9}$$

Coin toss example - a frequentist approach

- By the law of large numbers,

$$\hat{\sigma}_X^2 := \frac{\hat{\theta}_X(1-\hat{\theta}_X)}{n} \xrightarrow{\text{p}} \frac{\theta(1-\theta)}{n}. \tag{10}$$

- By the Slutsky's theorem,

$$\frac{\hat{\theta}_X - \theta}{\hat{\sigma}_X} \xrightarrow{\text{d}} \mathcal{N}(0,1), \tag{11}$$

and thus the Confidence Interval (CI) at level α ($100(1-\alpha)\%$ CI) is computed as

$$\Pr\Big(\hat{\theta}_X - Z_{1-\frac{\alpha}{2}}\hat{\sigma}_X < \theta < \hat{\theta}_X + Z_{1-\frac{\alpha}{2}}\hat{\sigma}_X\Big) \to 1-\alpha, \tag{12}$$

where Z_a is the inverse CDF of $\mathcal{N}(0,1)$.

Coin toss example - a frequentist approach

- Does this mean that the probability of θ being included in the CI is $1 - \alpha$?
    - No! θ is a fixed value. What's varying is the data X (and thus the CI computed from X).
    - So, the correct interpretation is, if we compute CIs for many datasets X generated from $p_{\text{true}}(x)$ over and over again, the fraction among those containing θ would approach $1 - \alpha$.
    - Does that sound intuitive?

Coin toss example - a Bayesian approach

- Believing that our model is true, we represent our uncertainty about θ as a prior distribution.

$$p(\theta) = \mathrm{Beta}(\theta; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}. \tag{13}$$

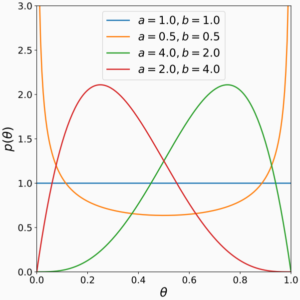

Figure 1: PDF of beta distribution with various parameters.

Coin toss example - a Bayesian approach

- Luckily, the posterior after observing X is still a Beta distribution, with parameters

$$p(\theta|X) = \mathrm{Beta}\left(\theta; a + \sum_{i=1}^{n} \mathbb{1}_{\{x_i = \mathsf{H}\}}, b + \sum_{i=1}^{n} \mathbb{1}_{\{x_i = \mathsf{T}\}}\right). \tag{14}$$

- The Credible Region (CR) $[L_X, U_X]$ at level α is defined as

$$\int_{L_X}^{U_X} p(\theta|X)\mathrm{d}\theta = 1 - \alpha. \tag{15}$$

- This requires a numerical approximation, but can directly be interpreted as, the probability of θ (after observing X) contained in $[L_X, U_X]$ is $1 - \alpha$!

Model selection for regression

- Assume we have a dataset $\mathcal{D} := (X, Y) = \{(x_i, y_i)\}_{i=1}^n$.
- We assume that $\mathcal{D}$ was generated from some function $y = f_\theta(x)$ plus some additive noise.

$$p(y|x;\theta) = \mathcal{N}(y|f_\theta(x), \sigma_y^2). \tag{16}$$

- What would be a proper form of $f_\theta(x)$?

$$f_\theta(x; \mathfrak{m}_1) = \theta_0 + \theta_1 x. \tag{17}$$

$$f_\theta(x; \mathfrak{m}_2) = \theta_0 + \theta_1 x + \theta_2 x^2. \tag{18}$$

Model selection for regression

- Is it right to compare the maximum likelihoods of models?

$$\max_{\theta} p(Y|X;\theta,\mathfrak{m}_1) < \max_{\theta} p(Y|X;\theta,\mathfrak{m}_2)? \tag{19}$$

- No, as you all might know, the infamous overfitting issue.

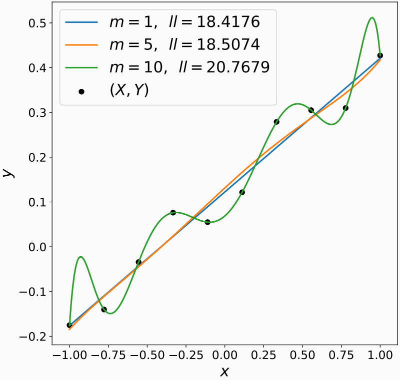

Figure 2: Maximum likelihood fits with various degrees.

Model selection for regression - frequentist approaches

- This happens because we don't take the model complexity into account.
- Akaike Information Criterion (AIC): penalize complex models (k : number of parameters).

$$\mathrm{AIC}(\mathfrak{m}) = 2k - \max_{\theta} \log p(Y|X; \theta, \mathfrak{m}). \tag{20}$$

- Alternative approaches: create more samples.
    - Cross-validation.
    - Bootstrapping.
- Also devise model-specific statistics whose distribution is well understood and easy to compute.

Model selection for regression - a Bayesian approach

- In Bayesian model, we can naturally define the marginal likelihood or evidence of data by averaging over all possible parameters.

$$p(Y|X;\mathfrak{m}) = \int p(Y|X,\theta;\mathfrak{m})p(\theta)\mathrm{d}\theta. \tag{21}$$

- We can even treat the model $\mathfrak{m}$ as a random variable, and compute the posterior probability of the model.

$$p(\mathfrak{m}|X,Y) = \frac{p(Y|X,\mathfrak{m})p(\mathfrak{m})}{p(Y|X)}. \tag{22}$$

Model selection for regression - a Bayesian approach

- To compare two models, we compute the Bayes factor.

$$\frac{p(Y|X, \mathfrak{m}_1)}{p(Y|X, \mathfrak{m}_2)}. \tag{23}$$

- Likewise, this requires a numerical approximation (sometimes given analytically though), but we can intuitively compare two different models without additional metrics, datasets (cross-validation), and model-specific statistics.

Frequentism vs Bayesianism - summary

- Frequentism
    - Probabilties are limiting frequencies.
    - Everything makes sense under the context of repeated experiments.
    - Model parameters are fixed.
    - In some sense, current data X is not that important!
    - Rather awkward definition of confidence interval, and requires some care for model comparison.
- Bayesianism
    - Probabilities are uncertainties.
    - Naturally defines uncertainties of parameters and even models via probabilities.
    - Intuitive definitions of confidence (credible region) and model comparison (Bayes factor).
    - Computations may be non-trivial.

Why I'm a Bayesian?

- In my opinion, it is more close to human way of learning concepts.
  - We have an initial knowledge, and it gets updated once we observe data.
  - Sequential update rule.

$$p(\theta|X_1, X_2) = \frac{p(X_2|\theta)p(\theta|X_1)}{p(X_2|X_1)} = \frac{p(X_2|\theta)p(X_1|\theta)p(\theta)}{p(X_2|X_1)P(X_1)}. \quad (24)$$

- Freedom to think of probabilities or uncertainties of non-trivial concepts (e.g., Bayesian nonparametric models)
- General principle for model assessment and comparison (I don't think the computation is the issue).
- It's cool.

Why uncertainty matters?

- Importance of knowing what you don't know.
  - Uber incident.
  - Racist algorithm by Google.
  - Medical diagnosis and decision making with financial data.
- Uncertainty guided sequential decision making
  - Bayesian optimization.
  - Reinforcement learning.
  - Active learning.

Elements of Bayesian learning

- We have a data X. We setup a model $\mathfrak{m}$ with a parameter θ.
- Inference: compute $p(\theta|X)$.

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}. \tag{25}$$

- Prediction: for a new data x_* and a function of interest f,

$$p(f(x_*)|X) = \int p(f(x_*)|\theta)p(\theta|X)\mathrm{d}\theta. \tag{26}$$

- Model comparision and criticism (posterior predictive checks, Bayes factors, ...).

The most useful identity

Monte-Carlo estimator of expectation.

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \xrightarrow{\mathrm{p}} \mathbb{E}_{p(x)}[f(x)] = \int f(x)p(x)\mathrm{d}x, \tag{27}$$

where $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} p(x)$.

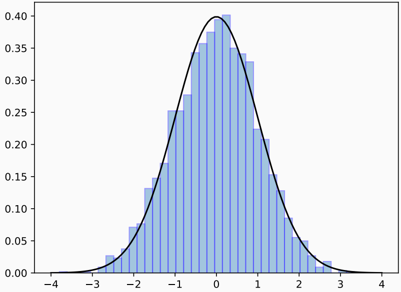

Figure 3: Monte-Carlo approximation for $\mathcal{N}(0, 1)$.

Recommended Readings

- *http://jakevdp.github.io/blog/2014/03/11/frequentism-and-bayesianism-a-practical-intro/*
- *http://www.stat.cmu.edu/~larry/=sml/Bayes.pdf*