Basics of Probabilistic/Bayesian Modeling and Parameter Estimation

Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

Jan 9, 2019
Some Announcements

- Prob-Stats refresher tutorial tomorrow (Thursday, Jan 10), 6:30pm-7:45pm, KD-101
  - Also posted some refresher slides on class webpage (under lecture-1 readings)
- A regular class this Saturday, Jan 12 (following Monday schedule)
- Sign up on Piazza if you haven’t already
- Regularly watch out for slides, readings etc., on class webpage
Probabilistic Modeling and Inference: The Fundamental Rules

- Keep in mind these two simple rules of probability: sum rule and product rule

\[ P(a) = \sum_b P(a, b) \quad \text{(Sum Rule)} \]

\[ P(a, b) = P(a)P(b|a) = P(b)P(a|b) \quad \text{(Product Rule)} \]

- Note: For continuous random variables, sum is replaced by integral: \( P(a) = \int P(a, b) \, db \)

- Another rule is the Bayes rule (can be easily obtained from the above two rules)

\[ P(b|a) = \frac{P(b)P(a|b)}{P(a)} = \frac{P(b)P(a|b)}{\int P(a, b) \, db} = \frac{P(b)P(a|b)}{\int P(b)P(a|b) \, db} \]

- All of probabilistic modeling and inference is based on consistently applying these two simple rules
Probabilistic Modeling

- Assume data $\mathbf{X} = \{ x_n \}_{n=1}^{N}$ generated from a probability distribution with parameters $\theta$

$$x_n \sim p(x|\theta, m), \quad n = 1, \ldots, N$$

- $p(x|\theta, m)$ is also known as the likelihood (a function of the parameters $\theta$)

- Assume a prior distribution $p(\theta|m)$ on the parameters $\theta$

- Note: Here $m$ collectively denotes “all other stuff” about the model, e.g.,
  - An “index” for the type of model being considered (e.g., “Gaussian”, “Student-t”, etc)
  - Any other (hyper)parameters of the likelihood/prior

- Note: Usually we will omit the explicit use of $m$ in the notation
  - In some situations (e.g., when doing model comparison/selection), we will use it explicitly

- Note: For some models, the likelihood is not defined explicitly using a probability distribution but implicitly via a probabilistic simulation process (more on such implicit probability models\(^\dagger\) later)

\(^\dagger\)Hierarchical Implicit Models and Likelihood-Free Variational Inference (Tran et al (NIPS 2017)
Probabilistic Modeling

- The prior distribution $p(\theta|m)$ plays a key role in probabilistic (especially Bayesian) modeling
  - Reflects our prior beliefs about possible parameter values before seeing the data

Can be “subjective” or “objective” (also a topic of debate, which we won’t get into)
- Subjective: Prior (our beliefs) derived from past experiments
- Objective: Prior represents “neutral knowledge” (e.g., uniform, vague prior)
- Can also be seen as a regularizer (connection with non-probabilistic view)

The goal of probabilistic modeling is usually one or more of the following
- Infer the unknowns/parameters $\theta$ given data $X$ (to summarize/understand the data)
- Use the inferred quantities to make predictions
Parameter Estimation/Inference

- Can infer the parameters by computing the posterior distribution (Bayesian inference)

\[
p(\theta|\mathbf{X}, m) = \frac{p(\mathbf{X}, \theta|m)}{p(\mathbf{X}|m)} = \frac{p(\mathbf{X}|\theta, m)p(\theta|m)}{\int p(\mathbf{X}|\theta, m)p(\theta|m) \, d\theta} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}
\]

- Note: **Marginal likelihood** \(p(\mathbf{X}|m)\) is another very important quantity (more on it later)

- Cheaper alternative: **Point Estimation** of the parameters. E.g.,
  
  - Maximum likelihood estimation (MLE): Find \(\theta\) that makes the observed data most probable
  \[
  \hat{\theta}_{ML} = \arg \max_{\theta} \log p(\mathbf{X}|\theta)
  \]
  
  - Maximum-a-Posteriori (MAP) estimation: Find \(\theta\) that has the largest posterior probability
  \[
  \hat{\theta}_{MAP} = \arg \max_{\theta} \log p(\theta|\mathbf{X}) = \arg \max_{\theta} [\log p(\mathbf{X}|\theta) + \log p(\theta)]
  \]
“Reading” the Posterior Distribution

- Posterior provides us a holistic view about $\theta$ given observed data
- A simple unimodal posterior distribution for a scalar parameter $\theta$ might look something like

![Posterior Distribution](image)

- Various types of estimates regarding $\theta$ can be obtained from the posterior, e.g.,
  - Mode of the posterior (same as the MAP estimate)
  - Mean and median of the posterior
  - Variance/spread of the posterior (uncertainty in our estimate of the parameters)
  - Any quantile (say $0 < \alpha < 1$ quantile) of the posterior, e.g., $\theta_\ast$ s.t. $p(\theta \leq \theta_\ast) = \alpha$
  - Various types of intervals/regions..
“Reading” the Posterior

- **100(1 − α)% Credible interval**: Region in which $1 − \alpha$ fraction of posterior’s mass resides

  $$C_\alpha(X) = (\ell, u) : p(\ell \leq \theta \leq u | X) = 1 - \alpha$$

- Credible Interval is not unique (there can be many 100(1 − α)% intervals)

- **Central Interval** is a symmetrized version of Credible Interval ($\alpha/2$ mass on each tail)

- Another useful interval: The $(1 - \alpha)$ Highest Probability Density (HPD) region is defined as

  $$C_\alpha(X) = \{\theta : p(\theta|X) \geq p^*\} \text{, s.t. } 1 - \alpha = \int_{\theta : p(\theta|X) \geq p^*} p(\theta|X)d\theta$$
“Reading” the Posterior

- CI, HPD, etc. can also be defined for multi-modal posteriors

Computing quantiles, CI, HPD, etc. may require inverting the CDF of the posterior
Using Posterior for Making Predictions

- Posterior can be used to compute the **posterior predictive distribution** (PPD) of new observation.
- The PPD of a new observation $x_*$ given previous observations is:

  $$p(x_*|X, m) = \int p(x_*|\theta, X, m) p(\theta|X, m) d\theta$$

  $$= \int p(x_*|\theta, m) p(\theta|X, m) d\theta$$

- Note: In the above, we assume that the observations are i.i.d. given $\theta$.
- Computing PPD requires doing a **posterior-weighted averaging** over all values of $\theta$.
- If the integral in PPD is intractable, we can approximate the PPD by **plug-in predictive**

  $$p(x_*|X, m) \approx p(x_*|\hat{\theta}, m)$$

  .. where $\hat{\theta}$ is a point estimate of $\theta$ (e.g., MLE/MAP).
- The plug-in predictive is the same as PPD with $p(\theta|X, m)$ approximated by a point mass at $\hat{\theta}$. 
Marginal Likelihood

- Recall the Bayes rule for computing the posterior

\[
p(\theta|X, m) = \frac{p(X, \theta|m)}{p(X|m)} = \frac{p(X|\theta, m)p(\theta|m)}{\int p(X|\theta, m)p(\theta|m)d\theta} = \text{Likelihood} \times \text{Prior} / \text{Marginal likelihood}
\]

- The denominator in the Bayes rule is the marginal likelihood (a.k.a. “model evidence”)

- Note that \(p(X|m) = \mathbb{E}_{p(\theta|m)}[p(X|\theta, m)]\) is the average/expected likelihood under model m

- For a good model, we would expect this “averaged” quantity to be large (most \(\theta\)'s will be good)

- Note that marginal likelihood is also like a “prior predictive distribution”
Model Selection and Model Averaging

- Marginal likelihood is hard-to-compute (due to integral) but a very useful quantity
- It can be used for doing model selection
  - Choose model $m$ that has largest posterior probability
    \[
    \hat{m} = \arg \max_m p(m|X) = \arg \max_m \frac{p(X|m)p(m)}{p(X)} = \arg \max_m p(X|m)p(m)
    \]
  - If all models are equally likely a priori then $\hat{m} = \arg \max_m p(X|m)$
  - If $m$ is a hyperparam, then $\arg \max_m p(X|m)$ is MLE-II based hyperparameter estimation
- Marginal likelihood can be used to compute $p(m|X)$ and then perform Bayesian Model Averaging
  \[
  p(x^*|X) = \sum_{m=1}^{M} p(x^*|X, m)p(m|X)
  \]
- BMA does a “double” averaging to make prediction since $p(x^*|X, m) = \int p(x^*|\theta, m)p(\theta|X, m)d\theta$
A Simple Parameter Estimation Problem

(for a single-parameter model)
(hyperparameter if any will be assumed to be fixed/known)
MLE via a simple example

Consider a sequence of $N$ coin tosses (call head = 0, tail = 1)

The $n^{th}$ outcome $x_n$ is a binary random variable $\in \{0, 1\}

Assume $\theta$ to be probability of a head (parameter we wish to estimate)

Each likelihood term $p(x_n | \theta)$ is Bernoulli: $p(x_n | \theta) = \theta^{x_n}(1 - \theta)^{1-x_n}$

Log-likelihood: $\sum_{n=1}^{N} \log p(x_n | \theta) = \sum_{n=1}^{N} x_n \log \theta + (1 - x_n) \log(1 - \theta)$

Taking derivative of the log-likelihood w.r.t. $\theta$, and setting it to zero gives

$$\hat{\theta}_{MLE} = \frac{\sum_{n=1}^{N} x_n}{N}$$

$\hat{\theta}_{MLE}$ in this example is simply the fraction of heads!

MLE doesn’t have a way to express our prior belief about $\theta$. Can be problematic especially when the number of observations is very small (e.g., suppose very few or zero heads when $N$ is small).
MAP via a simple example

- MAP estimation can incorporate a prior $p(\theta)$ on $\theta$
- Since $\theta \in (0, 1)$, one possibility can be to assume a Beta prior
  \[
  p(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1}
  \]
- $\alpha, \beta$ are called hyperparameters of the prior (these can have intuitive meaning; we'll see shortly)

Note that each likelihood term is still a Bernoulli: $p(x_n|\theta) = \theta^{x_n}(1 - \theta)^{1-x_n}$
MAP via a simple example (contd.)

- The log posterior probability = \( \sum_{n=1}^{N} \log p(x_n|\theta) + \log p(\theta) \)
- Ignoring the constants w.r.t. \( \theta \), the log posterior probability:
  \[
  \sum_{n=1}^{N} \{x_n \log \theta + (1 - x_n) \log(1 - \theta)\} + (\alpha - 1) \log \theta + (\beta - 1) \log(1 - \theta)
  \]
- Taking derivative w.r.t. \( \theta \) and setting to zero gives
  \[
  \hat{\theta}_{MAP} = \frac{\sum_{n=1}^{N} x_n + \alpha - 1}{N + \alpha + \beta - 2}
  \]

Note: For \( \alpha = 1, \beta = 1 \), i.e., \( p(\theta) = \text{Beta}(1, 1) \) (equivalent to a uniform prior), \( \hat{\theta}_{MAP} = \hat{\theta}_{MLE} \)

What hyperparameters represent intuitively? Hyperparameters of the prior (in this case \( \alpha, \beta \)) can often be thought of as “pseudo-observations”.

- \( \alpha - 1, \beta - 1 \) are the expected numbers of heads and tails, respectively, before seeing any data.
Recall that each likelihood term was Bernoulli: $p(x_n|\theta) = \theta^{x_n}(1-\theta)^{1-x_n}$

Let’s again choose the prior $p(\theta)$ as Beta: $p(\theta) = \text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$

The posterior distribution will be proportional to the product of likelihood and prior $p(\theta|X) \propto \prod_{n=1}^{N} p(x_n|\theta)p(\theta)$

$\propto \theta^{\alpha+\sum_{n=1}^{N}x_n-1}(1-\theta)^{\beta+N-\sum_{n=1}^{N}x_n-1}$

From simple inspection, note that the posterior $p(\theta|X) = \text{Beta}(\alpha + \sum_{n=1}^{N} x_n, \beta + N - \sum_{n=1}^{N} x_n)$

Here, finding the posterior boiled down to simply “multiply, add stuff, and identify the distribution”

Note: Can verify (exercise) that the normalization constant $= \frac{\Gamma(\alpha+\sum_{n=1}^{N} x_n)\Gamma(\beta+N-\sum_{n=1}^{N} x_n)}{\Gamma(\alpha+\beta+N)}$

To verify, make use of the fact that $\int p(\theta|X)d\theta = 1$

Here, the posterior has the same form as the prior (both Beta): property of conjugate priors.
Conjugate Priors

- Many pairs of distributions are conjugate to each other. E.g.,
  - Bernoulli (likelihood) + Beta (prior) ⇒ Beta posterior
  - Binomial (likelihood) + Beta (prior) ⇒ Beta posterior
  - Multinomial (likelihood) + Dirichlet (prior) ⇒ Dirichlet posterior
  - Poisson (likelihood) + Gamma (prior) ⇒ Gamma posterior
  - Gaussian (likelihood) + Gaussian (prior) ⇒ Gaussian posterior
  - and many other such pairs ..

- Easy to identify if two distributions are conjugate to each other: their functional forms are similar
  - E.g., recall the forms of Bernoulli and Beta
    \[
    \text{Bernoulli} \propto \theta^x(1 - \theta)^{1-x}, \quad \text{Beta} \propto \theta^{\alpha-1}(1 - \theta)^{\beta-1}
    \]

- More on conjugate priors when we look at exponential family distributions
Making Predictions

- Let's say we want to compute the probability that the next outcome \( x_{N+1} \in \{0, 1\} \) will be a head.

- The plug-in predictive distribution using a point estimate \( \hat{\theta} \) (e.g., using MLE/MAP)

\[
p(x_{N+1} = 1 | X) \approx p(x_{N+1} = 1 | \hat{\theta}) = \hat{\theta}
\]

or equivalently

\[
p(x_{N+1} | X) \approx \text{Bernoulli}(x_{N+1} | \hat{\theta})
\]

- The posterior predictive distribution (averaging over all \( \theta \) weighted by their posterior probabilities):

\[
p(x_{N+1} = 1 | X) = \int_0^1 p(x_{N+1} = 1 | \theta)p(\theta | X)d\theta
\]

\[
= \int_0^1 \theta \times \text{Beta}(\theta | \alpha + N_1, \beta + N_0)d\theta
\]

\[
= \mathbb{E}[\theta | X]
\]

\[
= \frac{\alpha + N_1}{\alpha + \beta + N}
\]

- Therefore the posterior predictive distribution:

\[
p(x_{N+1} | X) = \text{Bernoulli}(x_{N+1} | \mathbb{E}[\theta | X])
\]
Another Example: Estimating Gaussian Mean

- Consider $N$ i.i.d. observations $X = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \propto \exp\left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right]$$

$$p(X|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

- Assume the mean $\mu \in \mathbb{R}$ of the Gaussian is unknown and assume variance $\sigma^2$ to be known/fixed

- We wish to estimate the unknown $\mu$ given the data $X$

- Let’s do fully Bayesian inference for $\mu$ (not MLE/MAP)

- We first need a prior distribution for the unknown param. $\mu$

- Let’s choose a Gaussian prior on $\mu$, i.e., $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$ with $\mu_0, \sigma_0^2$ as fixed

- Therefore this is also a single-parameter model (only $\mu$ is the unknown)
Another Example: Estimating Gaussian Mean

The posterior distribution for the unknown mean parameter $\mu$

$$p(\mu|X) = \frac{p(X|\mu)p(\mu)}{p(X)} \propto \prod_{n=1}^{N} \exp \left[ -\frac{(x_n - \mu)^2}{2\sigma^2} \right] \times \exp \left[ -\frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right]$$

(Verify) The above posterior turns out to be another Gaussian $p(\mu|X) = \mathcal{N}(\mu|\mu_N, \sigma^2_N)$ where

$$\frac{1}{\sigma^2_N} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \bar{x} \quad \text{(where } \bar{x} = \frac{\sum_{n=1}^{N} x_n}{N} \text{)}$$

Making prediction: The posterior predictive distribution for a new observation $x_*$ will be

$$p(x_*|X) = \int p(x_*|\mu)p(\mu|X)d\mu = \int \mathcal{N}(x_*|\mu, \sigma^2)\mathcal{N}(\mu|\mu_N, \sigma^2_N)d\mu = \mathcal{N}(x_*|\mu_N, \sigma^2_N + \sigma^2)$$

Note that, in contrast, the plug-in predictive posterior, given a point estimate $\hat{\mu}$ would be

$$p(x_*|X) = \int p(x_*|\mu)p(\mu|X)d\mu \approx p(x_*|\hat{\mu}) = \mathcal{N}(x_*|\hat{\mu}, \sigma^2)$$