Basics of Probabilistic Modeling and Inference, Single Parameter Models

Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

Jan 9, 2018

Probabilistic Modeling and Inference

- Assume data $\boldsymbol{y} = \{y_1, \ldots, y_N\}$ generated from a probabilistic model (call it m) with parameters θ

$$y_1, \ldots, y_N \sim p(y|\theta, m)$$

- The Bayesian approach infers the unknowns θ by computing their posterior distribution

$$p(\theta|\boldsymbol{y}, m) = \frac{p(\boldsymbol{y}, \theta|m)}{p(\boldsymbol{y}|m)} = \frac{p(\boldsymbol{y}|\theta, m)p(\theta|m)}{\int p(\boldsymbol{y}|\theta, m)p(\theta|m)d\theta} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$

- Note: Here m is simply an "index" to refer to the model. For example, each m could refer to
    - A degree-k ($k = 0, 1, 2, \ldots$) polynomial (probabilistic) model for regression
- Note: Sometimes we will omit the explicit use of model index m in the notation
    - In some situations (e.g., when doing model comparison/selection), we will use it explicitly
- Note: The notion of what "model" refers to can be sometimes be more subtle (e.g., in hierarchical models, each distict value of a hyperparam would give rise to a different model). More on this later

Meaning of various terms..

- Let's again look at the Bayes rule for inferring the posterior distribution

$$p(\theta|\boldsymbol{y}, m) = \frac{p(\boldsymbol{y}|\theta, m)p(\theta|m)}{\int p(\boldsymbol{y}|\theta, m)p(\theta|m)d\theta} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}} \propto \text{Likelihood} \times \text{Prior}$$

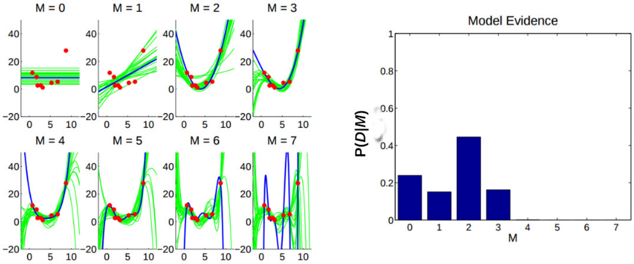
- Likelihood function $p(\boldsymbol{y}|\theta, m)$ or the "observation model" specifies how data is generated
  - It is also the probability of the observed data, given θ
- Prior distribution $p(\theta|m)$ specifies how likely different parameter values are *a priori*
  - As we'll see later, using a prior also corresponds to imposing a "regularizer" over θ
- Marginal likelihood $p(\boldsymbol{y}|m)$ is the average probability of the observed data $\boldsymbol{y}$ under model m

$$p(\boldsymbol{y}|m) = \int p(\boldsymbol{y}, \theta|m)d\theta = \int p(\boldsymbol{y}|\theta, m)p(\theta|m)d\theta$$

.. a very important quantity as we'll see later

More on Marginal Likelihood..

- The marginal likelihood is also called "model evidence". Recall its definition
$$p(\boldsymbol{y}|m) = \int p(\boldsymbol{y}, \theta|m)d\theta = \int p(\boldsymbol{y}|\theta, m)p(\theta|m)d\theta = \mathbb{E}_{p(\theta|m)}[p(\boldsymbol{y}|\theta, m)]$$
- It's the average probability of $\boldsymbol{y}$ for randomly drawn θ's from the model's prior $p(\theta|m)$
- We use the marginal likelihood as a reasonable notion of "goodness" of the model m



- Why: Because, for a good model, several parameters (rather than a select few) will fit the data "reasonably" well. Such a model is less likely to overfit and thus generalize better to future data
  - Caveat: The choice of prior $p(\theta|m)$ is important if using $p(\boldsymbol{y}|m)$ to do model selection

Making Predictions using Posterior Predictive Distribution

- In probabilistic modeling, making predictions requires computing the predictive distribution $p(\boldsymbol{y}_*|\boldsymbol{y}, m)$, i.e., probability distribution of new data $\boldsymbol{y}_*$, given past data $\boldsymbol{y}$
- This is formally defined by the so-called posterior predictive distribution

$$p(\boldsymbol{y}_*|\boldsymbol{y}, m) = \int p(\boldsymbol{y}_*, \theta|\boldsymbol{y}, m)d\theta \quad = \quad \int p(\boldsymbol{y}_*|\theta, \boldsymbol{y}, m)p(\theta|\boldsymbol{y}, m)d\theta$$
$$= \int p(\boldsymbol{y}_*|\theta, m)p(\theta|\boldsymbol{y}, m)d\theta$$

- This is basically the likelihood on new data with posterior-weighted averaging over all values of θ
- If posterior predictive is expensive to compute, we can approximate it by plug-in predictive

$$p(\boldsymbol{y}_*|\boldsymbol{y}, m) \approx p(\boldsymbol{y}_*|\hat{\theta}, m)$$

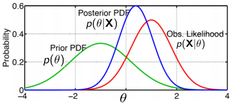
.. where $\hat{\theta}$ is a point estimate of θ (e.g., MLE/MAP)

- The marginal likelihood $p(\boldsymbol{y}|m)$ is a special case of posterior predictive (sort of a "prior predictive")
  - Reason: Recall that $p(\boldsymbol{y}|m) = \int p(\boldsymbol{y}|\theta, m)p(\theta|m)d\theta$

Estimating Parameters via Point Estimation

- Recall the definition of the posterior distribution over parameters (omitting the model index m)

$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$



- Although, typically the goal is to infer the posterior, sometimes we only want a point estimate
- Point Estimation finds the **single "best" estimate** of the parameters via optimization. E.g.,
  - Maximum likelihood estimation (MLE)

$$\hat{\theta} = \arg\max_{\theta} \log p(\mathbf{X}|\theta)$$

  - Maximum-a-Posteriori (MAP) estimation

$$\hat{\theta} = \arg\max_{\theta} \log p(\theta|\mathbf{X}) = \arg\max_{\theta} [\log p(\mathbf{X}|\theta) + \log p(\theta)]$$

- Point estimates doesn't provide us the uncertaintly in our estimate of θ

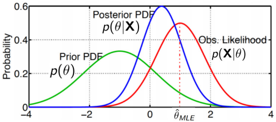
Point Estimation via MLE

- MLE finds the parameter θ that maximizes the (log-) likelihood $p(\mathbf{X}|\theta)$

$$\mathcal{L}(\theta) = \log p(\mathbf{X}|\theta) = \log p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \mid \theta)$$

- If the observations are i.i.d., $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \mid \theta) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta)$
- Maximum Likelihood parameter estimation

$$\boxed{\hat{\theta}_{MLE} = \arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta)}$$



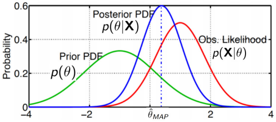
Point Estimation via MAP

- MAP estimation finds the parameter θ that maximizes the (log-) posterior probability $p(\theta|\mathbf{X})$

$$\mathcal{L}(\theta) = \log p(\theta|\mathbf{X}) = \log \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})}$$

- Again assuming i.i.d. observations, and noting that $p(\mathbf{X})$ is independent of θ

$$\boxed{\hat{\theta}_{MAP} = \arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta) + \log p(\theta)}$$



- Note: When the prior is uniform, MAP and MLE solutions are identical
- Despite using the prior, MAP is NOT considered a Bayesian approach (still gives a point estimate)

Point Estimation (MLE/MAP) vs Loss Function Minimization

- Recall the maximum Likelihood parameter estimation procedure

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \theta)$$

- We can also think of it as minimizing the **negative** log-likelihood (NLL)

$$\boxed{\hat{\theta}_{MLE} = \arg\min_{\theta} NLL(\theta)}$$

where $NLL(\theta) = -\sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \theta)$ is called the negative log-likelihood

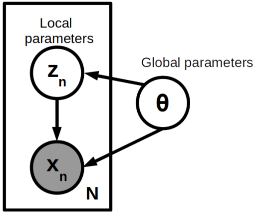
- Likewise, MAP parameter estimation can be shown to have the following form

$$\boxed{\hat{\theta}_{MAP} = \arg\min_{\theta} NLL(\theta) - \log p(\theta)}$$

- Can think of the NLL as a loss function and $-\log p(\theta)$ as a regularizer on θ
- Thus MLE is like empirical loss/risk minization (ERM) and MAP is like regularized ERM

“Hybrid” Inference

- Often we want to do point estimation for some parameters and fully Bayesian inference for others
- The choice depends on various factors (which we'll see later). But as a rule of thumb:
  - Perform fully Bayesian inference for “local variables”
  - Perform point estimation for “global” variables



- Local variables are data-point specific (so there is little data available to infer them)
- Global variables are shared by all data points (so usually plenty of data to infer them)

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 $\boldsymbol{x}_n$ is a binary random variable $\in \{0, 1\}$
- Assume θ to be probability of a head (parameter we wish to estimate)
- Each likelihood term $p(\boldsymbol{x}_n \mid \theta)$ is Bernoulli: $p(\boldsymbol{x}_n \mid \theta) = \theta^{\boldsymbol{x}_n}(1-\theta)^{1-\boldsymbol{x}_n}$
- Log-likelihood: $\sum_{n=1}^N \log p(\boldsymbol{x}_n \mid \theta) = \sum_{n=1}^N \boldsymbol{x}_n \log\theta + (1-\boldsymbol{x}_n)\log(1-\theta)$
- Taking derivative of the log-likelihood w.r.t. θ, and setting it to zero gives

$$\hat{\theta}_{MLE} = \frac{\sum_{n=1}^N \boldsymbol{x}_n}{N}$$

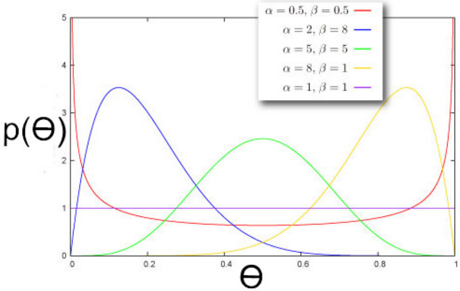
- $\hat{\theta}_{MLE}$ in this example is simply the fraction of heads!
- MLE doesn't have a way to express our prior belief about θ. Can be problematic especially when the number of observations is very small (e.g., suppose very few or zero heads when N is small).

MAP via a simple example

- MAP estimation can incorporate a prior $p(\theta)$ on θ
- Since $\theta \in (0,1)$, one possibility can be to assume a Beta prior

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- α, β are called hyperparameters of the prior (these can have intuitive meaning; we'll see shortly)



- Note that each likelihood term is still a Bernoulli: $p(\boldsymbol{x}_n|\theta) = \theta^{\boldsymbol{x}_n}(1-\theta)^{1-\boldsymbol{x}_n}$

MAP via a simple example (contd.)

- The log posterior probability $= \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\theta) + \log p(\theta)$
- Ignoring the constants w.r.t. θ, the log posterior probability:

$$\sum_{n=1}^{N} \{\boldsymbol{x}_n \log\theta + (1-\boldsymbol{x}_n)\log(1-\theta)\} + (\alpha-1)\log\theta + (\beta-1)\log(1-\theta)$$

- Taking derivative w.r.t. θ and setting to zero gives

$$\hat{\theta}_{MAP} = \frac{\sum_{n=1}^{N} \boldsymbol{x}_n + \alpha - 1}{N + \alpha + \beta - 2}$$

- 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 α, β) 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

Full Bayesian Inference via a simple example

- Recall that each likelihood term was Bernoulli: $p(\boldsymbol{x}_n|\theta) = \theta^{\boldsymbol{x}_n}(1-\theta)^{1-\boldsymbol{x}_n}$
- 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

$$
\begin{aligned}
p(\theta|\mathbf{X}) &\propto \prod_{n=1}^{N} p(\boldsymbol{x}_n|\theta)p(\theta) \\
&\propto \theta^{\alpha+\sum_{n=1}^{N}\boldsymbol{x}_n-1}(1-\theta)^{\beta+N-\sum_{n=1}^{N}\boldsymbol{x}_n-1}
\end{aligned}
$$

- From simple inspection, note that the posterior $p(\theta|\mathbf{X}) = \text{Beta}(\alpha+\sum_{n=1}^{N}\boldsymbol{x}_n, \beta+N-\sum_{n=1}^{N}\boldsymbol{x}_n)$
- Here, finding the posterior boiled down to simply "multipy, add stuff, and identify the distribution"
- Note: Can verify (exercise) that the normalization constant $= \frac{\Gamma(\alpha+\sum_{n=1}^{N}\boldsymbol{x}_n)\Gamma(\beta+N-\sum_{n=1}^{N}\boldsymbol{x}_n)}{\Gamma(\alpha+\beta+N)}$
    - To verify, make use of the fact that $\int p(\theta|\mathbf{X})d\theta = 1$
- 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) $\Rightarrow$ Beta posterior
  - Binomial (likelihood) + Beta (prior) $\Rightarrow$ Beta posterior
  - Multinomial (likelihood) + Dirichlet (prior) $\Rightarrow$ Dirichlet posterior
  - Poisson (likelihood) + Gamma (prior) $\Rightarrow$ Gamma posterior
  - Gaussian (likelihood) + Gaussian (prior) $\Rightarrow$ 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 exponental family distributions

Making Predictions

- Let's say we want to compute the probability that the next outcome $\boldsymbol{x}_{N+1} \in \{0, 1\}$ will be a head
- The plug-in predictive distribution using a point estimate $\hat{\theta}$ (e.g., using MLE/MAP)

$$p(\boldsymbol{x}_{N+1} = 1|\mathbf{X}) \approx p(\boldsymbol{x}_{N+1}|\hat{\theta}) = \hat{\theta} \qquad \underline{\text{or equivalently}} \qquad p(\boldsymbol{x}_{N+1}|\mathbf{X}) \approx \text{Bernoulli}(\boldsymbol{x}_{N+1} \mid \hat{\theta})$$

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

$$\begin{aligned} p(\boldsymbol{x}_{N+1} = 1|\mathbf{X}) &= \int_0^1 P(\boldsymbol{x}_{N+1} = 1|\theta)p(\theta|\mathbf{X})d\theta \\ &= \int_0^1 \theta \times \text{Beta}(\theta|\alpha + N_1, \beta + N_0)d\theta \\ &= \mathbb{E}[\theta|\mathbf{X}] \\ &= \frac{\alpha + N_1}{\alpha + \beta + N} \end{aligned}$$

- Therefore the posterior predictive distribution: $p(\boldsymbol{x}_{N+1}|\mathbf{X}) = \text{Bernoulli}(\boldsymbol{x}_{N+1} \mid \mathbb{E}[\theta|\mathbf{X}])$

Another Example: Estimating Gaussian Mean

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

$$
\begin{aligned}
p(x_n|\mu, \sigma^2) &= \mathcal{N}(x|\mu, \sigma^2) \propto \exp\left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right] \\
p(\mathbf{X}|\mu, \sigma^2) &= \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)
\end{aligned}
$$

- Assume the mean $\mu \in \mathbb{R}$ of the Gaussian is unknown and assume variance σ^2 to be known/fixed
- We wish to estimate the unknown μ given the data $\mathbf{X}$
- Let's do fully Bayesian inference for μ (not MLE/MAP)
- We first need a prior distribution for the unknown param. μ
- Let's choose a Gaussian prior on μ, i.e., $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$ with μ_0, σ_0^2 as fixed
- Therefore this is also a single-parameter model (only μ is the unknown)

Another Example: Estimating Gaussian Mean

- The posterior distribution for the unknown mean parameter μ

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

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

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

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

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

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

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

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

- Question: What happens when N is very large?