Probabilistic Modeling - An Illustration via Gaussian Mean Estimation

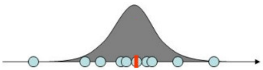
Piyush Rai

CS771 (Supplementary Notes/Slides)

August 14, 2018

A Toy Problem: Estimating the mean of a Gaussian

- Consider data consisting of N scalar-valued observations $x_1, \dots, x_N$
- Assume each observation is drawn i.i.d. from a one-dimensional Gaussian $\mathcal{N}(\mu, \sigma^2)$



- Would like to estimate the mean μ (assume that we know σ^2)
- One approach is to define an appropriate "loss function" and minimize it w.r.t. μ
- A possible loss function could be the sum of squared deviations from the mean

$$\mathcal{L}(\mu) = \sum_{n=1}^{N} (x_n - \mu)^2$$

- Minimizing it w.r.t. μ gives $\hat{\mu} = \frac{\sum_{n=1}^{N} x_n}{N}$ (i.e., the empirical mean of data)
- Can we solve this problem using a probabilistic approach?

The Probabilistic Approach

- Let's write down the probability of the N Gaussian-distributed observations (assumed i.i.d.)

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

- Note: The quantity $p(X|\mu)$ is also known as the likelihood
- Let's define the optimal μ as one that maximizes $p(X|\mu)$

$$\hat{\mu} = \arg\max_{\mu} p(X|\mu) = \arg\max_{\mu} \log p(X|\mu) = \arg\min_{\mu} \sum_{n=1}^{N}(x_n-\mu)^2$$

- The above procedure is commonly known as maximum likelihood estimation (MLE)
- The optimal μ will be the same as the previous loss function based approach, i.e., $\hat{\mu} = \frac{\sum_{n=1}^{N} x_n}{N}$
- MLE basically gave us the same solution. So what did we gain? **Stay tuned :-)**

Adding Prior Knowledge

- What if someone told us that μ is close to μ_0?
- Can add a "regularizer" $(\mu - \mu_0)^2$ to the objective function, and the solution would be

$$\hat{\mu} = \arg\min_{\mu} \left[\sum_{n=1}^{N} (x_n - \mu)^2 + (\mu - \mu_0)^2 \right] = \frac{\sum_{n=1}^{N} x_n + \mu_0}{N+1}$$

- Note that our estimate of μ has "shifted" a bit towards μ_0
- Question: What happens to our estimate when N is very large?
- Rather than adding a regularizer in ad-hoc way, can we do it in a more formal way?
- Yes. Using a "prior distribution" on μ

Prior Distribution

- Let's assume we have a probabilistic prior belief as to what μ might be (before seeing the data)
- Let us assume our belief is modeled by a Gaussian prior distribution on μ

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]$$

- The prior tells us that a prior we believe μ to be close to μ_0 with a "spread" σ_0^2
- Note: Gaussian prior not necessary; can use other distributions. But Gaussian has some benefits (e.g., computational ease; also makes sense in general in some cases)
- How do we now "update" our prior belief in the light of observed data X?
- To do this we need to combine the prior distribution $p(\mu)$ with the likelihood $p(X|\mu)$

Combining Prior and Likelihood..

- Enters the **Bayes rule**. Can define the posterior distribution of μ as

$$p(\mu|X) = \frac{p(X|\mu)p(\mu)}{p(X)} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal probability}}$$

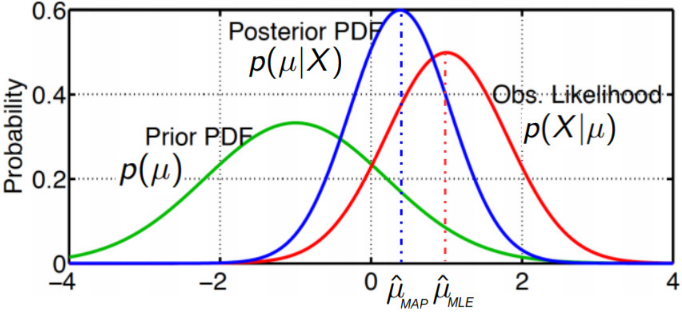
- We can find an optimal μ by maximizing the posterior distribution $p(\mu|X)$ w.r.t. μ

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

- The above procedure is commonly known as maximum-a-posteriori (MAP) estimation
- Plugging in $p(X|\mu)$ and $p(\mu)$ and simplifying, we get

$$\hat{\mu}_{MAP} = \arg\min_{\mu} \left[\sum_{n=1}^{N} \frac{(x_n - \mu)^2}{2\sigma^2} + \frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right] = \frac{\sum_{n=1}^{N} \frac{x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

MLE vs MAP: A Pictorial View



The Full Posterior

- MLE and MAP both only gave us a single best estimate of μ (also called a point estimate)
- However, we may sometimes be interested in the full posterior distribution over μ

$$p(\mu|X) = \frac{p(X|\mu)p(\mu)}{p(X)} = \frac{p(X|\mu)p(\mu)}{\int p(X|\mu)p(\mu)d\mu}$$

- The full posterior distribution provides a more complete picture about μ
- However, it is usually a hard problem since the integral to compute $p(X)$ is not always easy
- In some cases however (e.g., Gaussian mean estimation), the posterior can be computed easily

$$p(\mu|X) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$$

where

$$\mu_N = \frac{\sum_{n=1}^N \frac{x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}} \quad \text{and} \quad \sigma_N^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}} \qquad \text{(exercise: verify)}$$

- Note that the posterior is the same distribution as the prior - both are Gaussian (this happens when likelihood and prior are conjugate to each other)

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

$$\text{Bernoulli} \propto \theta^{x}(1-\theta)^{1-x}, \quad \text{Beta} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Making Predictions

- So we have estimated μ, either via MLE/MAP or its full posterior distribution
- Suppose, for a new observation x_*, we want to compute its predictive distribution $p(x_*|X)$
- This too can be done in two ways
  - Compute the plug-in predictive distribution using the MLE/MAP point estimate $\hat{\mu}$

$$p(x_*|X) = \int p(x_*, \mu|X)d\mu = \int p(x_*|\mu, X)p(\mu|X)d\mu \approx p(x_*|\hat{\mu}, X) = \underbrace{p(x_*|\hat{\mu})}_{\text{since data is i.i.d.}}$$

  - Compute the posterior predictive distribution by averaging over the posterior of μ

$$p(x_*|X) \quad = \quad \int p(x_*, \mu|X)d\mu = \int p(x_*|\mu, X)p(\mu|X)d\mu = \int p(x_*|\mu)p(\mu|X)d\mu$$

- Posterior averaged prediction is more robust (and also more informative)
  - Caveat: In general, much harder to compute as compared to the plug-in prediction but can be done in closed form in this case since $p(x_*|\mu)$ and $p(\mu|X)$ both are Gaussians