Parameter Estimation in Probabilistic Models: An Example

CS772A: Probabilistic Machine Learning

Piyush Rai

Estimating a Coin's Bias: MLE

- Consider a sequence of N coin toss outcomes (observations)
- Each observation y_n is a binary random variable. Head: $y_n = 1$, Tail: $y_n = 0$
- Each y_n is assumed generated by a **Bernoulli distribution** with param $\theta \in (0,1)$

Probability of a head

Likelihood or observation model

$$p(y_n|\theta) = \text{Bernoulli}(y_n|\theta) = \theta^{y_n}(1-\theta)^{1-y_n}$$

- Here θ the unknown param (probability of head). Want to estimate it using MLE

assuming i.i.d. data

- Log-likelihood: $\sum_{n=1}^{N} \log p(y_n|\theta) = \sum_{n=1}^{N} [y_n \log\theta + (1-y_n)\log(1-\theta)]$
- Maximizing log-lik, or minimizing neg. log-lik (NLL) w.r.t. θ gives

I tossed a coin 5 times – gave 1 head and 4 tails. Does it means θ = 0.2?? The MLE approach says so. What is I see 0 head and 5 tails. Does it mean θ = 0?

$$\theta_{MLE} = \frac{\sum_{n=1}^{N} y_n}{N}$$

Thus MLE solution is simply the fraction of heads! ☺ Makes intuitive sense!

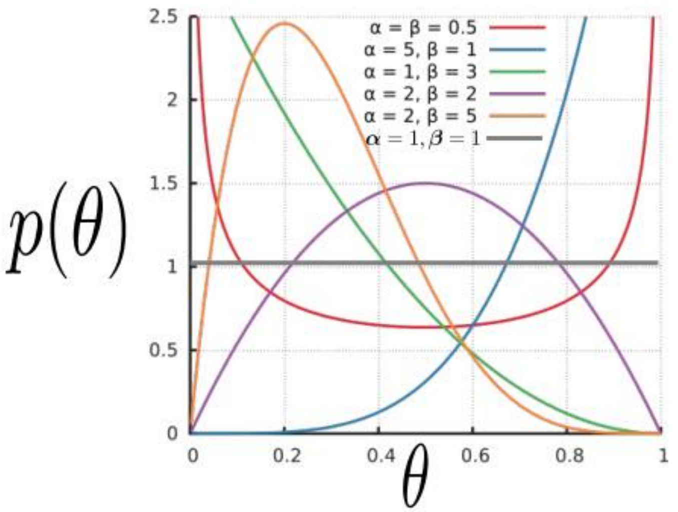
Indeed, with a small number of training observations, MLE may overfit and may not be reliable. An alternative is MAP estimation which can incorporate a prior distribution over θ

Estimating a Coin's Bias: MAP

- Let's again consider the coin-toss problem (estimating the bias of the coin)
- Each likelihood term is Bernoulli

$$p(y_n|\theta) = \text{Bernoulli}(y_n|\theta) = \theta^{y_n}(1-\theta)^{1-y_n}$$

- Also need a prior since we want to do MAP estimation
- Since $\theta \in (0,1)$, a reasonable choice of prior for θ would be Beta distribution



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

The gamma function

α and β (both non-negative reals) are the two hyperparameters of this Beta prior

Using $\alpha = 1$ and $\beta = 1$ will make the Beta prior a uniform prior

Can set these based on intuition, cross-validation, or even learn them

Estimating a Coin's Bias: MAP

- The log posterior for the coin-toss model is log-lik + log-prior

$$LP(\theta) = \sum_{n=1}^{N} \log p(y_n|\theta) + \log p(\theta|\alpha,\beta)$$

- Plugging in the expressions for Bernoulli and Beta and ignoring any terms that don't depend on θ, the log posterior simplifies to

$$LP(\theta) = \sum_{n=1}^{N} [y_n \log\theta + (1-y_n)\log(1-\theta)] + (\alpha-1)\log\theta + (\beta-1)\log(1-\theta)$$

- Maximizing the above log post. (or min. of its negative) w.r.t. θ gives

$$\theta_{MAP} = \frac{\sum_{n=1}^{N} y_n + \alpha - 1}{N + \alpha + \beta - 2}$$

Using $\alpha = 1$ and $\beta = 1$ gives us the same solution as MLE

Recall that $\alpha = 1$ and $\beta = 1$ for Beta distribution is in fact equivalent to a uniform prior (hence making MAP equivalent to MLE)

Prior's hyperparameters have an interesting interpretation. Can think of $\alpha - 1$ and $\beta - 1$ as the number of heads and tails, respectively, before starting the coin-toss experiment (akin to "pseudo-observations")

Such interpretations of prior's hyperparameters as being "pseudo-observations" exist for various other prior distributions as well (in particular, distributions belonging to "exponential family" of distributions

Estimating a Coin's Bias: Fully Bayesian Inference

- In fully Bayesian inference, we compute the posterior distribution
- Bernoulli likelihood: $p(y_n|\theta) = \text{Bernoulli}(y_n|\theta) = \theta^{y_n}(1-\theta)^{1-y_n}$
- Beta prior: $p(\theta) = \text{Beta}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- The posterior can be computed as

Number of heads (N_1)

Number of tails (N_0)

$\theta^{\sum_{n=1}^N y_n}(1-\theta)^{N-\sum_{n=1}^N y_n}$

$$p(\theta|\boldsymbol{y}) = \frac{p(\theta)p(\boldsymbol{y}|\theta)}{p(\boldsymbol{y})} = \frac{p(\theta)\prod_{n=1}^N p(y_n|\theta)}{p(\boldsymbol{y})} = \frac{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\prod_{n=1}^N \theta^{y_n}(1-\theta)^{1-y_n}}{\int \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\prod_{n=1}^N \theta^{y_n}(1-\theta)^{1-y_n}d\theta}$$

- Here, even without computing the denominator (marg lik), we can identify the posterior
  - It is Beta distribution since $p(\theta|\boldsymbol{y}) \propto \theta^{\alpha+N_1-1}(1-\theta)^{\beta+N_0-1}$
  - Thus $p(\theta|\boldsymbol{y}) = \text{Beta}(\theta|\alpha+N_1,\beta+N_0)$

Hint: Use the fact that the posterior must integrate to 1 $\int p(\theta|\boldsymbol{y})d\theta = 1$

Exercise: Show that the normalization constant equals $\frac{\Gamma(\alpha+\sum_{n=1}^N x_n)\Gamma(\beta+N-\sum_{n=1}^N x_n)}{\Gamma(\alpha+\beta+N)}$

- Here, finding the posterior boiled down to simply "multiply, add stuff, and identify"
- Here, posterior has the same form as prior (both Beta): property of **conjugate priors**

Conjugacy and Conjugate Priors

- Many pairs of distributions are conjugate to each other
  - 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 ..

Not true in general, but in some cases (e.g., the variance of the Gaussian likelihood is fixed)

- Tip: If two distr are conjugate to each other, their functional forms are similar
  - Example: Bernoulli and Beta have the forms

$$\text{Bernoulli}(y|\theta) = \theta^{y}\,(1-\theta)^{1-y}$$

$$\text{Beta}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

This is why, when we multiply them while computing the posterior, the exponents get added and we get the same form for the posterior as the prior but with just updated hyperparameter. Also, we can identify the posterior and its hyperparameters simply by inspection

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

Making Predictions

- Suppose we want to compute the prob that the next outcome x_{N+1} will be head (=1)
- The plug-in predictive distribution using a point estimate $\hat{\theta}$ (e.g., using MLE/MAP)

$$p(x_{N+1} = 1|\mathbf{X}) \approx p(x_{N+1} = 1|\hat{\theta}) = \hat{\theta}$$ or equivalently $$p(x_{N+1}|\mathbf{X}) \approx \text{Bernoulli}(x_{N+1} \mid \hat{\theta})$$

- The posterior predictive distribution (averaging over all θ's weighted by their respective posterior probabilities)

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

Expectation of θ w.r.t. the Beta posterior distribution

- Therefore the PPD is $p(x_{N+1}|\mathbf{X}) = \text{Bernoulli}(x_{N+1} \mid \mathbb{E}[\theta|\mathbf{X}])$