Bayesian methods

Cong Ma

Outline

- Review ridge regression and Lasso
- Bayesian methods and MAP
- Understand regularization from MAP perspective

The Bayesian Paradigm

Given a parameter θ, we assume observations are generated according to $p(z|\theta)$. In our work so far, we have treated the parameter θ like a fixed, deterministic, but unknown quantity while the observation z is the realization of a random process.

We will now consider **probabilistic models** for θ in addition to our data.

- This allows us to incorporate **prior information** we have about θ (i.e. information about likely values of θ we have *before collecting any data*).
- It also allows us to make statements about our **confidence** in different estimates of θ.

Example: Unfair coin

Suppose you toss a single coin 6 times and each time it comes up "heads." It might be reasonable to say that we are 98% sure that the coin is unfair, biased towards heads.

Formally, we can think about this in a hypothesis testing framework using a binomial probabilistic model. Let $z :=$ number of "heads".

$$\text{hypothesis: } \mathbb{P}(\text{heads}) \equiv \theta > 0.5$$

$$p(z|\theta) = \binom{n}{z} \theta^z (1-\theta)^{n-z}$$

$$p(\theta > 0.5|z) = ?$$

Example: (cont.)

The problem with this is that

$$p(\theta \in H_0|x)$$

implies that θ is a random, not deterministic, quantity.

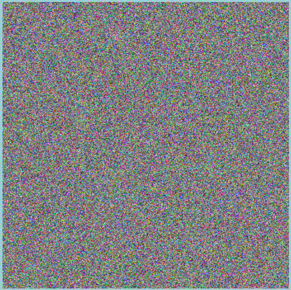
So, while "confidence" statements are very reasonable and in fact a normal part of "everyday thinking," this idea can not be supported from the classical perspective.

All of these "deficiencies" can be circumvented by a change in how we view the parameter θ.

Example: Image processing

In many imaging problems, we have a good sense of what "natural" images should look like.

Likely



Unlikely

This prior information can be exploited to improve image denoising, deblurring, reconstruction, and analysis.

Bayes Rule

If we view θ as the realization of a random variable with density $p(\theta)$, then we can work with the generative (or forward) model

$$\underbrace{p(\theta)}_{\text{prior}} \rightarrow \theta^* \rightarrow \underbrace{p(z|\theta^*)}_{\text{likelihood}} \rightarrow z.$$

We are interested in the inverse problem

$$z \rightarrow p(\theta|z) \rightarrow \hat{\theta}.$$

Bayes Rule (Bayes, 1763) shows that

$$p(\theta|z) = \frac{p(z|\theta)\, p(\theta)}{p(z)} = \frac{p(z|\theta)\, p(\theta)}{\int p(z|\widetilde{\theta})\, p(\widetilde{\theta})\, d\widetilde{\theta}}$$

Once we can compute this posterior distribution, confidence measures such as $p(\theta \in H_0|z)$ are perfectly legitimate quantities to ask for.

Example: Coin toss

Suppose you toss a single coin 6 times and each time it comes up "heads." Mathematically, we can model the problem as follows. Let $\theta = \mathbb{P}(\text{Heads})$. The data (the number of heads z in $n = 6$ tosses) follows a binomial distribution $p(z|\theta) = \binom{n}{z}\theta^z(1-\theta)^{n-z}$. The mathematical equivalent of the question "is the coin probably biased" is the probability $\mathbb{P}(\theta > 0.5|z = 6)$.

Suppose we assume $p(\theta) = \mathrm{Unif}(0,1)$ (all values of θ are equally probable before we begin to flip the coin, and $\mathbb{P}(\theta > \frac{1}{2}) = \frac{1}{2}$). Now compute

$$p(\theta|z) = \frac{p(z|\theta)p(\theta)}{\int p(z|\widetilde{\theta})p(\widetilde{\theta})d\widetilde{\theta}} = \frac{\theta^6}{\int \widetilde{\theta}^6 d\widetilde{\theta}} = \frac{\theta^6}{\frac{1}{7}\widetilde{\theta}^7|_0^1} = 7\,\theta^6\ .$$

Then

$$\mathbb{P}\left(\theta > \frac{1}{2}\,|\,z = 6\right) = \int_{\frac{1}{2}}^{1} 7\widetilde{\theta}^6 d\widetilde{\theta} = \widetilde{\theta}^7|_{\frac{1}{2}}^{1} = 1 - 2^{-7} = 0.984\ .$$

(If we chose a different prior we would get a different answer!)

Bayesian statistical models

Definition: Bayesian statistical model

A Bayesian statistical model is composed of a *data generation model*, $p(z|\theta)$, and a *prior* distribution on the parameters, $p(\theta)$.

The prior distribution (or "prior" for short) models the uncertainty in the parameter. More specifically, $p(\theta)$ models our knowledge - or a lack thereof - prior to collecting data.

Notice that

$$p(\theta|z) = \frac{p(z|\theta)\, p(\theta)}{p(z)} \propto p(z|\theta)\, p(\theta)$$

Hence, $p(\theta|z)$ is proportional to the likelihood function multiplied by the prior.

Elements of Bayesian Analysis

(a) Joint distribution

$$p(z,\theta) = p(z|\theta)p(\theta)$$

(b) Marginal distributions

$$p(z) = \int p(z|\theta)p(\theta)d\theta$$

$$p(\theta) = \int p(z|\theta)p(\theta)dz \text{ (“prior”)}$$

(c) Posterior distribution

$$p(\theta|z) = \frac{p(z,\theta)}{p(z)} = \frac{p(z|\theta)p(\theta)}{\int p(z|\widetilde{\theta})p(\widetilde{\theta})d\widetilde{\theta}}$$

Maximum A posteriori

Definition

Maximum A Posteriori (MAP) estimator - the value of θ where $p(\theta|z)$ is maximized:

$$\widehat{\theta}_{\mathrm{MAP}}(z) = \arg\max_{\widetilde{\theta}} p(\widetilde{\theta}|z) = \arg\max_{\widetilde{\theta}} p(z|\widetilde{\theta})p(\widetilde{\theta})$$

Example: Binomial + Beta

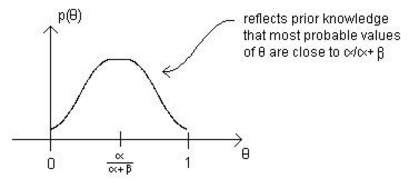
$$p(z|\theta) = \binom{n}{z}\theta^z(1-\theta)^{n-z}, 0 \leq \theta \leq 1$$

= binomial likelihood

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

= Beta prior distribution

where $\Gamma(\alpha) = \int_0^\infty z^{\alpha-1}e^{-z}dz$ is the Gamma function



Example: (cont.)

- Joint Density

$$p(z,\theta) = \left[\frac{\binom{n}{z}}{B(\alpha,\beta)}\right]\theta^{\alpha+z-1}(1-\theta)^{n-z+\beta-1}$$

- Marginal Density

$$p(z) = \left[\binom{n}{z}\frac{1}{B(\alpha,\beta)}\right]B(\alpha+z,\beta+n-z)$$

- Posterior Density

$$p(\theta|z) = \underbrace{\frac{\theta^{\alpha+z-1}(1-\theta)^{\beta+n-z-1}}{B(\alpha+z,\beta+n-z)}}_{\substack{\text{beta density with parameters} \\ \alpha'=\alpha+z \\ \beta'=\beta+n-z}}$$

Linear regression with prior

$$y \mid x = w^\top x + \varepsilon;$$
$$w \sim \mathcal{N}(0, r^2 \boldsymbol{I})$$

How to compute MAP for w?

Ridge vs. LASSO

$$\hat{\theta}_{\text{Ridge}} = \arg\min_{\theta} \left\{ \frac{1}{2}\|y - X\theta\|_2^2 + \frac{\sigma_\epsilon^2}{2\sigma_\theta^2}\|\theta\|_2^2 \right\}$$
$$\hat{\theta}_{\text{LASSO}} = \arg\min_{\theta} \left\{ \frac{1}{2}\|y - X\theta\|_2^2 + \frac{\sigma_\epsilon^2 \lambda}{2}\|\theta\|_1 \right\}$$

In both cases, we attempt to find a θ which (a) is a good fit to our data and (b) adheres to prior information captured by either the ℓ_2 or ℓ_1 norm of θ.

When should we use one vs. the other?

In general, the LASSO estimator favors *sparser* θ – i.e., θ with more zero-valued elements. There is no closed-form expression for the LASSO estimate.

Overview

The multivariate Gaussian linear model...

- ... with a multivariate Gaussian prior $\Longrightarrow$ **ridge regression**
- ... with a multivariate Laplace prior $\Longrightarrow$ **LASSO (least absolute shrinkage and selection operator) regression**

These models and methods appear in a wide variety of modern machine learning settings.