Machine Learning

MAP Estimation

Dr. Mehran Safayani

safayani@iut.ac.ir

safayani.iut.ac.ir

https://www.aparat.com/mehran.safayani

https://github.com/safayani/machine_learning_course

Department of Electrical and computer engineering, Isfahan university of technology, Isfahan, Iran

Maximum a posteriori (MAP) Estimation

- MLE Recall:

In Maximum Likelihood Estimation (MLE), we used iid samples $\mathbf{x} = (x_1, \ldots, x_n)$ from some distribution with unknown parameter(s) θ, in order to estimate θ.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \; L(\mathbf{x} \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{n} f_X(x_i \mid \theta)$$

You might have been thinking: shouldn't we be trying to maximize "$\mathbb{P}(\theta \mid x)$" instead? Well, this doesn't make sense **unless Θ is a R.V.**! And this is where Maximum A Posteriori (MAP) Estimation comes in.

https://courses.cs.washington.edu/courses/cse312/20su/files/student_drive/7.5.pdf

- Now, we are in the Bayesian framework, meaning that our unknown

parameter is a random variable θ

- This means, we will have some belief distribution $p(\theta)$(think of this as a density function over all possible values of the parameter), and after observing data x, we will have a new/updated belief distribution $p(\theta|x)$.
- Using Bayes theorem:

Bayes theorem

$$\underbrace{P(\theta|X)}_{a\ posterior\ probability} = \frac{\overbrace{P(X|\theta)}^{Likelihood}\ \overbrace{P(\theta)}^{a\ prior\ probability}}{\underbrace{P(X)}_{Marginal\ probability}} \approx P(X|\theta)P(\theta)$$

MAP estimation Example

- We have observed data X $= \{x_1, x_2, \dots, x_n\}$
  and we want to estimate the mean μ of a normal distribution. We assume a norm
  al prior on μ with mean μ_0 and variance
  τ^2 , and we assume the data X is drawn from a normal distribution with mean μ
  and known variance σ2.

$$P(X|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

$$P(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\tau^2}\right)$$

$$P(\mu|X) \propto P(X|\mu) \cdot P(\mu)$$

MAP estimation

$$P(\mu|X) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\cdot\exp\left(-\frac{(\mu-\mu_0)^2}{2\tau^2}\right)$$

$$\ln P(\mu|X) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{(\mu-\mu_0)^2}{2\tau^2} + \text{constant}$$

- Set the derivative to zero

$$\frac{d}{d\mu}\ln P(\mu \mid X) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) + \frac{\mu_0-\mu}{\tau^2} = 0$$

MAP estimation

$$\sum_{i=1}^{n} x_i - n\mu + \frac{\sigma^2}{\tau^2}(\mu_0 - \mu) = 0$$

$$n\mu + \frac{\sigma^2}{\tau^2}\mu = \sum_{i=1}^{n} x_i + \frac{\sigma^2}{\tau^2}\mu_0$$

$$\mu(n + \frac{\sigma^2}{\tau^2}) = \sum_{i=1}^{n} x_i + \frac{\sigma^2}{\tau^2}\mu_0$$

$$\mu = \frac{\sum_{i=1}^{n} x_i + \frac{\sigma^2}{\tau^2}\mu_0}{n + \frac{\sigma^2}{\tau^2}}$$

Regularization

L_2_regularization:

$$\underset{w}{\text{Min}} \frac{1}{2N} \sum_{n=1}^{N} [y_n - x_n^T w]^2 + \lambda \|w\|_2$$

$$\nabla L(w) = \frac{-1}{N} x^T (\text{y - xw})$$

$$+$$

$$\nabla \Omega(w) = 2\lambda \text{w}$$

Prove the solution for w:

$$w_{ridge}^* = (x^T x + \lambda' I)^{-1} x^T \text{y} \quad , \frac{\lambda'}{2N} = \lambda$$

نکته: ماتریس $x^T x + \lambda' I$ معکوس پذیر است.

Regression as MLE estimator: Recall

$$
\begin{aligned}
\mathbf{w}_{\text{lse}} &\overset{(a)}{=} \arg\min_{\mathbf{w}} -\log p(\mathbf{y}, \mathbf{X} \mid \mathbf{w}) \\
&\overset{(b)}{=} \arg\min_{\mathbf{w}} -\log p(\mathbf{X} \mid \mathbf{w}) p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) \\
&\overset{(c)}{=} \arg\min_{\mathbf{w}} -\log p(\mathbf{X}) p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) \\
&\overset{(d)}{=} \arg\min_{\mathbf{w}} -\log p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) \\
&\overset{(e)}{=} \arg\min_{\mathbf{w}} -\log \left[\prod_{n=1}^{N} p(y_n \mid \mathbf{x}_n, \mathbf{w}) \right]
\end{aligned}
$$

Regression as MLE estimator: Recall

$$\stackrel{(f)}{=} \arg\min_{\mathbf{w}} -\log\left[\prod_{n=1}^{N} \mathcal{N}(y_n \mid \mathbf{x}_n^\top \mathbf{w}, \sigma^2)\right]$$

$$= \arg\min_{\mathbf{w}} -\log\left[\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_n - \mathbf{x}_n^\top \mathbf{w})^2}\right]$$

$$= \arg\min_{\mathbf{w}} -N\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum_{n=1}^{N} \frac{1}{2\sigma^2}(y_n - \mathbf{x}_n^\top \mathbf{w})^2$$

$$= \arg\min_{\mathbf{w}} \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \mathbf{x}_n^\top \mathbf{w})^2$$

Ridge Regression as MAP estimator

فرض می کنیم که اجزای $\mathbf{w}$ دارای توزیع گاوسین مستقل و یکنواخت با میانگین صفر و واریانس $\frac{1}{\lambda}$ است.

$$\mathbf{w}_{\text{ridge}} = \arg\min_{\mathbf{w}} -\log p(\mathbf{w} \mid \mathbf{X},\mathbf{y})$$

$$\overset{(a)}{=} \arg\min_{\mathbf{w}} -\log \frac{p(\mathbf{y},\mathbf{X}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{y},\mathbf{X})}$$

$$\overset{(b)}{=} \arg\min_{\mathbf{w}} -\log p(\mathbf{y},\mathbf{X} \mid \mathbf{w})p(\mathbf{w})$$

$$\overset{(c)}{=} \arg\min_{\mathbf{w}} -\log p(\mathbf{y} \mid \mathbf{X},\mathbf{w})p(\mathbf{w})$$

•

Ridge Regression as MAP estimator

$$= \arg\min_{\mathbf{w}} -\log\left[p(\mathbf{w}) \prod_{n=1}^{N} p(y_n \mid \mathbf{x}_n, \mathbf{w})\right]$$

- $$= \arg\min_{\mathbf{w}} -\log\left[\mathcal{N}\left(\mathbf{w} \mid 0, \frac{1}{\lambda}\mathbf{I}\right) \prod_{n=1}^{N} \mathcal{N}(y_n \mid \mathbf{x}_n^\top \mathbf{w}, \sigma^2)\right]$$

$$= \arg\min_{\mathbf{w}} -\log\left[\frac{1}{\left(2\pi\frac{1}{\lambda}\right)^{D/2}} e^{-\frac{\lambda}{2}\|\mathbf{w}\|^2} \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left(y_n - \mathbf{x}_n^\top \mathbf{w}\right)^2}\right]$$

$$= \arg\min_{\mathbf{w}} \sum_{n=1}^{N} \frac{1}{2\sigma^2}(y_n - \mathbf{x}_n^\top \mathbf{w})^2 + \frac{\lambda}{2} \| \mathbf{w} \|^2.$$