Lecture 5: Regression (contd.)

24 January 2024 11:02

Gradient Descent

- Initialisation : $w \leftarrow w_0$
- Repeat until convergence : $\| \nabla_w E \| < \varepsilon$

$$w_{t+1} \leftarrow w_t - \eta \nabla_w E \qquad \eta : \text{learning rate}$$

$$E(w, D) = \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

$$\nabla_w E = \sum_{i=1}^{n} \nabla_w E_i$$

Gradient descent is excellent in terms of accuracy but expensive in computation

Stochastic Gradient Descent

update state: $w_{t+1} \leftarrow w_t - \nabla_w E(w, x_i, y_i)$, ($i$ is randomly chosen) $\rightarrow \nabla_w \left((w^T x_i - y_i)^2 \right)$

Faster computation than Gradient descent

$$E(w, (x_i, y_i)_{i \in B})$$

$$B \subseteq [n] = \{1, 2, \ldots n\}$$

update:

$$w_{t+1} \leftarrow w_t - \eta \sum_{i \in B} \nabla E_i$$

MLE (recap) :

Let $D = \{(x_i, y_i)_{i \in [n]}\}$ and θ be parameter to be estimated.

We are interested in $\underset{\theta}{\arg\max}\ P(D|\theta) = \theta_{MLE}$ ($P(D|\theta)$: likelihood function)

Recall the coin toss example here.

In the coin toss, $\theta_{MLE} = \frac{1}{n} \sum_{j=1}^{n} y_j$

Now, what if we had a prior knowledge regarding θ?

Eg: In coin toss, we know a-priori that θ has a normal-like distribution

Can we update / incorporate this idea? This leads to the notion of Maximum Aposteriori Estimate.

Maximum Aposteriori Estimate (MAP) :

$P(\text{parameters} \mid \text{observation})$

θ cause ↑ — D effect ↑

Bayesian inference: (Prior, Likelihood)

$$P(\theta|D) = \frac{P(D|\theta)\,P(\theta)}{P(D)}$$

We call $P(\theta|D)$ posterior belief

$$\theta_{MAP} \in \arg\max_\theta P(\theta|D) = \arg\max_\theta \left[P(D|\theta)P(\theta)\right]$$

Now, $\log P(\theta|D) = \log P(D|\theta) + \log P(\theta)$

$$\Rightarrow \theta_{MAP} \in \arg\max_\theta \left[\log P(D|\theta) + \log P(\theta)\right]$$

If $P(\theta)$ is independent of θ (say uniform) then $\theta_{MAP} = \theta_{MLE}$
i.e., $P(\theta)$ is a constant.

Ex: Likelihood of observing k heads in n tosses.

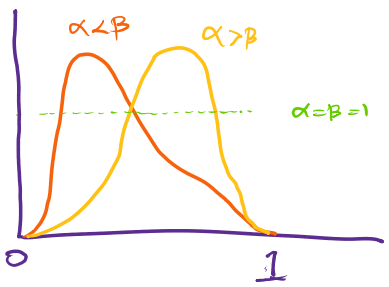
$$P(D|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \qquad \text{Bin}(n,\theta)$$

Here, $\theta_{MLE} = k/n$

Beta distribution

pmf: $P(\theta) = \frac{1}{c}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

If $\alpha = \beta = 1$ $P(\theta)$ becomes constant, we've a uniform distribution.



- Beta includes a large family of distributions in $[0,1]$
- Beta is a conjugate prior of binomial distribution.

Conjugate prior (C.P):
Let $P(D|\theta) \sim d_1$ $P(\theta) = d_2$
Then, $P(\theta)$ is a C.P of $P(D|\theta)$ if $P(\theta|D) \sim d_2$

$$P(\theta|D) \propto \underbrace{\theta^k(1-\theta)^{n-k}}_{P(D|\theta)}\ \underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{P(\theta)}$$

$$\propto \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} \sim \text{Beta}(k+\alpha,\ n-k+\beta)$$

$$\arg\max_\theta P(\theta|D) = \arg\max_\theta \log P(\theta|D)$$

$$= \text{const.} + (k+\alpha-1)\log\theta + (n-k+\beta-1)\log(1-\theta)$$

$$\frac{k+\alpha-1}{\quad} \qquad \left(\text{If } \alpha=\beta=1,\ \theta_{MAP} = \theta_{MLE}\right)$$

$= \text{const.} + (k+\alpha-1)\log\theta + (n-k+\beta-1)\log(1-\theta)$

$$\theta_{MAP} = \frac{k+\alpha-1}{n+\alpha+\beta-2} \quad (\theta_{MLE})$$

(If $\alpha=\beta=1$, $\theta_{MAP} = \theta_{MLE}$ → uniform → no prior info.)

Conjugate prior examples

1. Bernoulli / Binomial ↔ Beta
2. Geometric ↔ Beta
3. Categorical ↔ Dirichlet
4. Normal ↔ Normal

Conjugate prior for (univariate) Gaussian with known variance :-

likelihood $P(D|\theta) \sim N(\mu, \sigma^2)$, prior $P(\theta) \sim N(\mu_0, \sigma_0^2)$

$D = \{x_1, \dots x_n\}$

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

$$f(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{\frac{-1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}$$

$$P(\theta|D) \propto P(D|\theta)\,P(\theta) \propto \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}$$

What will be $\tilde{\sigma}^2, \tilde{\mu}$?

$$\propto \exp\left\{\frac{-1}{2\tilde{\sigma}^2}\sum(x_i-\tilde{\mu})^2\right\} \sim N(\tilde{\mu}, \tilde{\sigma}^2)$$

MAP estimate for linear regression

$$y_i = \omega^T x_i + \epsilon_i \sim N(\omega^T x_i, \sigma^2) \quad \text{where, } \epsilon_i \sim N(0, \sigma^2)$$

$$P(\theta|D) \propto \exp\left\{\frac{-1}{2\sigma^2}\sum(y_i-\omega^T x_i)^2\right\}$$

$$\theta_{MLE} = \arg\min_{\omega} \sum_{i=1}^{n}(y_i-\omega^T x_i)^2$$

Now, $P(\omega) \sim N\left(0, \frac{1}{\lambda}I\right)$

λ: hyperparameter.

$$P(\omega) = \frac{1}{(2\pi/\lambda)^{d/2}} \exp\left\{\frac{-\lambda}{2}\omega^T\omega\right\}$$

$$\propto \exp\left\{-\frac{\lambda}{2}\|\omega\|^2\right\}$$

Now, $P(\omega|D) \propto P(D|\omega)\,P(\omega)$

Multivariate Gaussian distribution:

For $\bar{x} \in \mathbb{R}^d$,

$$P(\bar{x}) \sim N(\bar{\mu}, \Sigma)$$

(Σ: covariance matrix, $\mathbb{R}^{d\times d}$)

where: $\Sigma[i][j] = \sigma_{ij} = \text{cov}(x_i, x_j)$

$$P(X) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

$$\arg\max_{w} \left(\log P(D|w) + \log P(w) \right) :$$

$$\arg\min_{w} \left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \frac{\lambda}{2} \|w\|^2 \right\} \equiv \left\{ \frac{1}{2\sigma^2} \|Xw - y\|^2 + \frac{\lambda}{2} \|w\|^2 \right\}$$

deterministic linear regression

Regularizer.