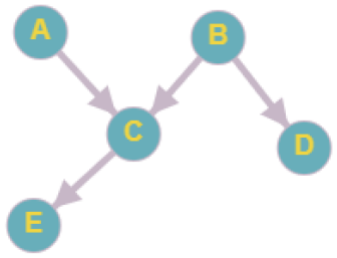
Homework

Alexander Chernyavskiy

November 2022

1 Chain rule



$$P(A, B, C, D, E) = P(E|C)P(C|A, B)P(D|B)P(A)P(B)$$

2 The Least Squares method

$$\exists w : \forall \{(x_i, y_i)\}_{i=1}^n \in D \to y_i = wx_i + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Assuming that $x_i \sim i.i.d$, we could write (for $1D$ case):

$$p(y|w, \sigma^2) = \prod_{i=1}^n p(y_i|w, \sigma^2) = \prod_{i=1}^n \mathcal{N}(y_i; wx_i, \sigma^2) =$$

$$= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y_i - wx_i)^2}{2\sigma^2}\right] = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - wx_i)^2\right]$$

$$\Rightarrow \log p(y|w, \sigma^2) = -\frac{n}{2}\left(\log(2\pi) + \log(\sigma^2)\right) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - wx_i)^2$$

As MLE $= \max_{w,\sigma} \log p(y|w, \sigma^2)$ and $-\log(\cdot)$ is a convex function, so we can just zero all derivatives:

$$\frac{\partial \log p(y|w,\sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i - wx_i)^2 = 0$$

$$\Rightarrow (\sigma^2)^* = \frac{1}{n}\sum_{i=1}^{n}(y_i - wx_i)^2$$

$$\frac{\partial \log p(y|w,\sigma^2)}{\partial w} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i x_i - wx_i^2) = 0$$

$$\Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n} y_i x_i - \frac{w}{\sigma^2}\sum_{i=1}^{n} x_i^2 = 0$$

$$\Rightarrow w^* = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Thus, MLE corresponds to finding a minimum of the function

$$\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - wx_i)^2|_{\sigma^2=const}.$$

If to consider linear regression in Bayesian approach, and posterior as normally distributed $p(w) \sim \mathcal{N}(0, \sigma_1^2 I)$, we have:

$$p(w|y,x) = \frac{p(y|w,x)p(w|x)}{p(y|x)}$$

Considering the fact that $p(y|x)$ does not depend on the parameter vector w, we can write

$$\log p(w|y,x) = \log p(y|w,x) + \log p(w|x) - \log p(y|x)$$

$$\Rightarrow w^* = \arg\max_w \{\log p(y|w,x) + \log p(w|x)\},$$

$$p(y|w,x) \sim \mathcal{N}(w^\top x, \sigma^2 I),$$

$$p(w|x) = p(w) \sim \mathcal{N}(0, \sigma_1^2 I),$$

Remember what the multivariate Normal distribution is:

$$\mathcal{N}(x;\mu,\Sigma) = \frac{\exp\left[-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right]}{\sqrt{(2\pi)^n \det \Sigma}},\ \Sigma > 0$$

Thus, the log-likelihood for $\log p(y|w,x)$ would be denoted as follows:

$$\log p(y|w,x) = -\frac{1}{2}(n\log(2\pi) + log(\det \Sigma)) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^\top \Sigma^{-1}(x_i - \mu) =$$

$$= -\frac{n}{2}(\log(2\pi) + log(\sigma^2)) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - w^\top x_i)^\top (x_i - w^\top x_i)$$

Analogously, for posterior we have,

$$\log p(w|x) = -\frac{n}{2}(\log(2\pi) + log(\sigma_1^2)) - \frac{1}{2\sigma_1^2}\sum_{i=1}^{n} w^\top w$$

If we need to find maximum w,

$$\log p(w^*|y,x) = \max_w \left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - w^\top x_i)^\top (x_i - w^\top x_i) - \frac{1}{2\sigma_1^2}\sum_{i=1}^{n} w^\top w \right] =$$

$$= \min_w \left[\sum_{i=1}^{n}(x_i - w^\top x_i)^2 + \underbrace{\frac{\sigma^2}{\sigma_1^2}}_{C^{-1}} \sum_{i=1}^{n} w^\top w \right]$$

Having σ^2 fixed as some constant, we derived Tikhonov regularisation – this model is also known as Ridge regression, so is there the function in `sklearn`[1].

3 A conjugate prior

Given the Bayes' formula:

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

If there is a parametric family of distributions Φ and the prior belongs to it, the prior is called conjugate to the likelihood iff posterior also belongs to Φ.

Assuming having a problem of estimation of *categorical distribution* with a PMF:

$$L(x|\theta) = \prod_{k=1}^{K} \theta^{[x=k]} = \prod_{k=1}^{K} \theta_k^{x_k}$$

It is necessary to show that *Dirichlet prior distribution* with a PDF

$$\alpha = (\alpha_1, \ldots, \alpha_K),$$

$$Dir(\theta_1, \ldots, \theta_K|\alpha) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

is the conjugate prior for the categorical distribution likelihood.

So, let us count the posterior:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta} = \frac{\prod_{i=1}^{n}\prod_{k=1}^{K}\theta_k^{x_{ki}}\theta_k^{\alpha_k - 1}}{\int \prod_{i=1}^{n}\prod_{k=1}^{K}\theta_k^{x_{ki}}\theta_k^{\alpha_k - 1}d\theta} =$$

$$= \frac{\prod_{k=1}^{K}\theta_k^{\sum_{i=1}^{n} x_{ki} + \alpha_k - 1}}{\int \prod_{k=1}^{K}\theta_k^{\sum_{i=1}^{n} x_{ki} + \alpha_k - 1}d\theta} = \left| c_k := \sum_{i=1}^{n} x_{ki} \right| = \frac{\prod_{k=1}^{K}\theta_k^{c_k + \alpha_k - 1}}{B(c+\alpha)} =$$

$$= Dir(\theta|\alpha + c)$$

[1] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html

This distribution could be chosen because this prior looks like a distribution on $(K-1)$-dimensional simplex. This relationship is used in Bayesian statistics to estimate the underlying parameter p of a categorical distribution given a collection of n samples. Intuitively, we can view the prior vector θ as pseudocounts, i.e. as representing the number of observations in each category that we have already seen. Then we simply add in the counts for all the new observations (the vector c) in order to derive the posterior distribution. [From Wikipedia]

4 Convergence rate

Let we have a sampled (i.i.d) $x_i{}_{i=1}^n$ where each element is normally distributed like $x_i \sim \mathcal{N}(w, 1)$ and $w \sim \mathcal{N}(0, 1)$.

Thus, we can have MLE of w:

$$\log p(x_i|w) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - w)^2,$$

$$\Rightarrow w^*_{MLE} = \arg\max_w \log p(x_i|w) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

All the x_i are i.i.d, and $\mathbb{E}x_i = w < \infty$, $\mathbb{D}x_i = 1 < \infty$, so writing the limit having considered that, according to the central limit theorem $\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} x_i - \mu) \to \mathcal{N}(0, 1)$:

$$\forall \varepsilon > 0 \lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i - w^*_{MLE}\right| > \varepsilon\right) = 0$$

Then $\forall \varepsilon > 0$,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i - w^*_{MLE}\right| \le \varepsilon\right) = 1 - P\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i - w^*_{MLE}\right| > \varepsilon\right) \ge$$

$$\ge \text{ (Chebyshev ineq.) } 1 - \frac{1}{\varepsilon^2 n},$$

so it converges as $n \to \infty$, and

$$\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n} x_i - w^*_{MLE}\right| \le \varepsilon, \Rightarrow \left|\frac{1}{n}\sum_{i=1}^{n} x_i - w^*_{MLE}\right| \le \frac{\varepsilon}{\sqrt{n}},$$

and convergence rate for MLE is $\mathcal{O}(\frac{1}{\sqrt{n}})$

On the other hand, we can have MAP of w:

$$\log p(w|x_i) = \log p(x_i|w) + \log p(w) - const =$$
$$= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - w)^2 - \frac{1}{2}\sum_{i=1}^{n} w^2 - const,$$
$$\Rightarrow w^*_{MAP} = \arg\max_w \log p(w|x_i) = \frac{1}{n+1}\sum_{i=1}^{n} x_i.$$

$$\forall \varepsilon > 0 \lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i - w^*_{MAP}\right| > \varepsilon\right) = 0.$$

A proof of convergence is analogous to aforementioned for the MLE. Given that, asymptotic convergence rate should be the same, $\mathcal{O}(\frac{1}{\sqrt{n}})$.

Next, let us estimate $|w^*_{MLE} - w^*_{MAP}|$ with triangle inequality:

$$\forall \varepsilon_1, \varepsilon_2 > 0 \; |w^*_{MLE} - w^*_{MAP}| \leq \left| -\frac{1}{n}\sum_{i=1}^{n} x_i + w^*_{MLE}\right| + \left|\frac{1}{n}\sum_{i=1}^{n} x_i - w^*_{MAP}\right| \leq \frac{\varepsilon_1 + \varepsilon_2}{\sqrt{n}},$$

Errata: this estimate could be in $\mathcal{O}(\frac{1}{n})$.

thus, MAP estimation converges to MLE as the sample n size grows, so that means that Bayesian approach and frequency approaches give the same result in case of any "large" sample for the normal prior (the prior distribution "vanishes") – this result could be known as Bernstein-von Mises Theorem.

I could be mistaken but this links helped me:

- link 1: Do Bayesian priors become irrelevant with large sample size?
- link 2: Why is the convergence rate important?
- link 3: Convergence Rates for the MAP of an Exponential Family and Stochastic Mirror Descent – an Open Problem

5 Profit (HAND-WAVY!)

Let us consider a multi-armed bandit problem[2]; a bandit is a tuple $\langle \mathcal{A}, \mathcal{R} \rangle$:

- $\mathcal{A}$ is a set of known m actions;
- $\mathcal{R}^a = \mathbb{P}[r|a]$ is an unknown distribution over rewards;
- at each step t the agent makes an action $a_t \in \mathcal{A}$, and the environment generates a reward $r_t \sim \mathcal{R}^{a_t}$;
- the goal of the agent is to maximise cumulative reward $\sum_{\tau=1}^{t} r_\tau$

[2] A David Silver's lecture

Denoting $Q(a) := \mathbb{E}[r|a]$ and the optimal value V^* as $V^* := \max_{a\in\mathcal{A}} Q(a)$, so the problem of maximising reward could be equivalent to minimise the total regret

$$L_t = \mathbb{E}\left[\sum_{\tau=1}^{t} V^* - Q(a_\tau)\right] = \sum_{a\in\mathcal{A}} \mathbb{E}[N_t(a)] \overbrace{(V^* - Q(a))}^{\text{a gap } \Delta_a} = \sum_{a\in\mathcal{A}} \mathbb{E}[N_t(a)]\Delta_a,$$

where $N_t(a)$ is a number of selections for action a at the moment t. A good algorithm ensures small counts for large gaps.

To ensure a sublinear solution for the exploration/exploitation dilemma, we have ε-greedy algorithm

$$a_t = \begin{cases} \arg\max_{a\in\mathcal{A}} \hat{Q}_t(a), & \text{with prob. } 1-\varepsilon, \\ U[1,m], & \text{with prob. } \varepsilon, \end{cases}$$

where

$$\hat{Q}_t(a) = \frac{1}{N_t(a)} \sum_{t=1}^{T} r_t \mathbb{I}_{a_t}(a).$$

Chernoff-Hoeffding bound. Let $\{x_i\}_{i=1}^n$ to be a sequence of random variables with common range $[0;1]$ s.t. $\mathbb{E}[x_t|x_1,\ldots,x_{t-1}] = \mu$, then for $S_n = \sum_{i=1} x_n$ and $\forall c \geq 0$

$$P\left(|S_n - \mu| \geq c\right) \leq e^{-\frac{2c^2}{n}}$$

Writing a probability of the a-th arm is chosen at time t like a sum of conditional probabilties P(chosen a at step t|explore)+P(chosen a at step t|non explore):

$$P(\text{chosen } a \text{ at step } t) = \frac{\varepsilon}{m} + \left(1 - \frac{\varepsilon}{m}\right) \arg\max_{a\in\mathcal{A}} \hat{Q}_t(a) \leq$$
$$\leq \frac{\varepsilon}{m} + \left(1 - \frac{\varepsilon}{m}\right) \underbrace{P(\hat{Q}_{t-1}(a) \geq \hat{Q}_{t-1}(a^*))}_{\text{a prob. to take a suboptimal action } a \neq a^*}$$

The arm could be chosen like that if the probability of taking the action is above the average or below the average in case of the optimal action:

$$\mu_a = \mathbb{E}[\hat{Q}(a_t)],$$
$$P(\hat{Q}_{t-1}(a) \geq \hat{Q}_{t-1}(a^*)) = P\left(\hat{Q}_{t-1}(a) \geq \mu_a + \frac{\Delta_a}{2}\right) + P\left((\hat{Q}_{t-1}(a^*) \leq \mu^* - \frac{\Delta_a}{2}\right) =$$
$$= P\left(\hat{Q}_{t-1}(a) - \mu_a \geq \frac{\Delta_a}{2}\right) + P\left(-\hat{Q}_{t-1}(a^*) + \mu^* \geq \frac{\Delta_a}{2}\right) \leq 2e^{-\frac{\delta^2}{2}N_t(a)}$$

It is a known result[3] that for those policies satisfying $\mathbb{E}N_t(a) = \Omega(\log n)$ the regret should be the best possible, so P(chosen a at step t) $= \Omega(e^{\log n}) = \Omega(n)$

[3] the most of results are here