ML - Exercise Sheet 5 - Kernels

Exercise 1 - logSumExp Trick

$P_* = \prod_{i=1}^{n} P_i$ $\quad P_+ = \sum_{i=1}^{N} P_i$

ⓐ Compute P_* and P_+ using P_is $\in [0,1]$ - why numerical instabilities:

P_*: when we work with probabilities $\in [0,1]$, if we multiply many probabilities to compute P_*, the result can become extremely tiny and may underflow to zero in the computer however the true result is not zero.

Example ⇒ let $n = 1000$ and $P_i = 0.01$ for $i = 1, \dots, n$

$$P_* = \prod_{i=1}^{1000} P_i = (0.01)^{1000} = (10^{-2})^{1000} = 10^{-2000} \Longrightarrow$$ which is very small to be considered as a numer not equal to zero in computer.

P_+: when we add numbers ~~with~~ from $[0,1]$, if we have very small numbers for sum of P_is these numners maybe loose their effect when comparing but relatively high numbers in this interval.

Example: $P_1 = 0.9 \quad P_2 = 0.05 \quad P_3 = (0.01)^{-10}$

when we add these numners in computer result ~~is~~ could be as:

$P_+ = \sum_1^3 P_i = P_1 + P_2 + P_3 = 0.9 + 0.05 + 0.01^{-10} = 0.95 \Rightarrow$ the effect of P_3 is not considered.

ⓑ $\ln P_*$ given $\ln p_i$'s $\Rightarrow P_i = [k_1, k_2, \dots, k_n]$ s.t $k_i = \ln p_i$

$$P_* = \prod_{i=1}^{n} p_i \Rightarrow \ln P_* = \ln\left(\prod_{i=1}^{n} p_i\right) = \ln\left(\prod_{i=1}^{n} e^{k_i}\right) = \sum_{i=1}^{n} \ln e^{k_i} = \sum_{i=1}^{n} k_i$$

if we are given the values $k_i = \ln p_i$ we could use liner sum of the given values. Since we don't comput the product of tiny numbers, this method avoids underflow.

ⓒ Comput $\ln P_+$ given $\ln p_i$'s: $\ln P_+ = \ln \sum_{i=1}^{n} \exp(\ln p_i)$ $\quad e^{\ln p_i - c} \cdot e^{c} = e^{\ln p_i}$ ✓

ⓘ logSumExp Trick:

$$\ln P_+ = \ln \sum_{i=1}^{n} \exp(\ln p_i) \xrightarrow{c = \max \ln p_i} \ln P_+ = \ln\left(\sum_{i=1}^{n} \exp(\ln p_i - c) \cdot \exp(c)\right)$$

$$= \ln(\exp(c)) + \ln \sum_{i=1}^{n} \exp(\ln p_i - c) = \underbrace{c + \ln \sum_{i=1}^{n} \exp(\ln p_i - c)}$$

The resut is the same with original formula, we just pull out the maximum value out of exponential.

ⓘⓘ why this method reduce numerical stability issues?

when we subtract the max $\ln p_i$ from all P_is the all terms $(\ln p_i - \max(\ln p_i))$ will be less or equal to zero and the exponentials will be less or equal to one always. Hence we will not have overflow for any term $(\ln p_i - c)$ and tiny probabilities ~~belong to [0,1]~~ will beome numbers in $[0,1]$ preventing underflow.

Exercise Sheet 5 - Kernels

Exercise 2 : Prove that Gussian Kernel is Mercer kernel

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$

$$\|x-y\|^2 = \|x\|^2 + \|y\|^2 - 2x^Ty$$

$$\Rightarrow K(x,y) = \exp\left(-\frac{\|x\|^2+\|y\|^2-2x^Ty}{2\sigma^2}\right) = \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)\cdot\exp\left(-\frac{\|y\|^2}{2\sigma^2}\right)\cdot\exp\left(\frac{x^Ty}{\sigma^2}\right)$$

Taylor Series for $\exp(x^Ty/\sigma^2)$:

$$\exp\left(\frac{x^Ty}{\sigma^2}\right) = \sum_{n=0}^{\infty}\frac{(x^Ty)^n}{n!\,\sigma^{2n}}$$

multinomial Theorem for $(x^Ty)^n$:

$$(x^Ty)^n = \left(\sum_{i=1}^{d} x_i y_i\right)^n = \sum_{k_1+k_2+\dots+k_d=n} \underbrace{\frac{n!}{k_1!k_2!\dots k_d!}}_{\text{constant}} \prod_{i=1}^{d}(\underbrace{x_i^{k_i}}_{\text{function of } x}\ \underbrace{y_i^{k_i}}_{\text{function of } y})$$

d → dimension = number of features)

$$\Rightarrow (x^Ty)^n = \sum_k f_k(x)\cdot f_k(y)$$

$$f_k(x) = \prod_{i=1}^{d}\frac{x_i^k}{k_1!k_2!\dots k_d!}$$

now if we consider $\exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)$ together with $f_k(x)$ and also $\exp\left(-\frac{\|y\|^2}{2\sigma^2}\right)$ with $f_k(y)$:

$$K(x,y) = \underbrace{\exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)\cdot\left(f_k(x)\right)}_{\text{function of } x} * \underbrace{\exp\left(-\frac{\|y\|^2}{2\sigma^2}\right)\cdot\left(f_k(y)\right)}_{\text{function of } y}$$

$$K(x,y) = \langle F(x), F(y)\rangle \quad \blacksquare$$

inner product in some other feature space