Assignment 2 for #70240413-0
"Statistical Machine Learning"

Kui XU, 2016311209

2017/03/23

1 Boosting: from Weak to Strong

Choose one problem from the 1.1 and 1.2. A bonus would be given if you finished the both.

1.1 Calculus

The gamma function is defined by (assuming x > 0)

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du. \tag{1}$$

(1) Prove that $\Gamma(x+1) = x\Gamma(x)$.

(2) Also show that

$$\int_0^1 u^{a-1}(1-u)^{b-1} du = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{2}$$

Solution: For Question (1), we can prove it by Using integration by parts, the steps are as follows:

$$\begin{aligned}
\Gamma(x+1) &= \int_0^\infty u^x e^{-u}\, du \\
&= \left[-u^x e^{-u}\right]_0^\infty + \int_0^\infty x u^{x-1} e^{-u}\, du \\
&= \lim_{u\to\infty}(-u^x e^{-u}) - (0e^{-0}) + x\int_0^\infty u^{x-1} e^{-u}\, du \\
&= x\int_0^\infty u^{x-1} e^{-u}\, du \\
&= x\Gamma(x)
\end{aligned} \tag{3}$$

As we know, when $u \to \infty$, $-u^x e^{-u} \to 0$, so the equation is proved.

Solution: For Question (2), we know that the left of the equation is a Beta function. From the definitions, we can express the equation which we want to prove as :

$$\Gamma(a+b)B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{4}$$

It's a double integral, the expansion formula is as follows:

$$\begin{aligned}
\Gamma(a+b)B(a,b) &= \int_0^\infty u^{a+b-1} e^{-u} du \int_0^1 v^{a-1}(1-v)^{b-1} dv \\
&= \int_0^\infty \int_0^1 (uv)^{a-1}[u(1-v)]^{b-1} u e^{-u} du\, dv
\end{aligned} \tag{5}$$

Then we do a transformation $w = uv$, $z = u(1-v)$. The inverse transformation is $u = w+z$, $v = w/(w+z)$, the corresponding ranges of them are $w \in (0, \infty)$ and $u \in (0, \infty)$.
The absolute value of the Jacobian is

$$\left| \nabla \frac{\partial(u, v)}{\partial(w, z)} \right| = \frac{1}{(w+z)} \tag{6}$$

Next, we use the changed of variables to do a double integral, the equation above becomes:

$$\begin{aligned} & \int_0^\infty \int_0^\infty w^{a-1} z^{b-1} (w+z) e^{-(w+z)} \frac{1}{w+z} dw\, dz \\ & = \int_0^\infty \int_0^\infty w^{a-1} z^{b-1} e^{-(w+z)} dw\, dz \\ & = \int_0^\infty w^{a-1} e^{-w} dw \int_0^\infty z^{b-1} e^{-z} dz \\ & = \Gamma(a)\Gamma(b) \end{aligned} \tag{7}$$

Finally the equation is proved.

1.2 Optimization

Use the Lagrange multiplier method to solve the following problem:

$$\begin{aligned} \min_{x_1, x_2} \quad & x_1^2 + x_2^2 - 1 \\ s.t. \quad & x_1 + x_2 - 1 = 0 \\ & 2x_1 - x_2 \geq 0 \end{aligned} \tag{8}$$

Solution: Consider the above equation is consist of inequality constraint functions and it is a nonlinear optimization problem, we can use the lagrange multiplier method with KKT condition to solve it. We construct the Lagrangian function for the problem:

$$\mathcal{L}(x, \lambda, \mu) = x_1^2 + x_2^2 - 1 + \lambda \cdot \Big(x_1 + x_2 - 1 \Big) + \mu \cdot \Big(2x_1 - x_2 \Big) \tag{9}$$

The certain conditions which are called KKT condition should satisfy,

$$\begin{aligned} \frac{\partial(\mathcal{L})}{\partial(X)} |_X = 0 \\ \lambda_j \neq 0 \\ \mu_k \geq 0 \\ \mu_k \cdot \Big(x_1^* + x_2^* - 1 \Big) = 0 \\ x_1^* + x_2^* - 1 = 0 \\ 2x_1^* - x_2^* \leq 0 \end{aligned} \tag{10}$$

We set up the equations:

$$\begin{aligned} \frac{\partial(\mathcal{L}, x, \lambda, \mu)}{\partial(x_1)} &= 2x_1 + \lambda + 2\mu = 0 \\ \frac{\partial(\mathcal{L}, x, \lambda, \mu)}{\partial(x_2)} &= 2x_2 + \lambda - \mu = 0 \\ \frac{\partial(\mathcal{L}, x, \lambda, \mu)}{\partial(\lambda)} &= x_1 + x_2 - 1 = 0 \\ \frac{\partial(\mathcal{L}, x, \lambda, \mu)}{\partial(\mu)} &= 2x_1 - x_2 = 0 \end{aligned} \tag{11}$$

We solve them:

$$\begin{aligned} x_1 &= \frac{1}{3} \\ x_2 &= \frac{2}{3} \\ \lambda &= \frac{2}{9} \\ \mu &= -\frac{10}{9} \end{aligned} \tag{12}$$

Choose one problem from the following 1.3 and 1.4. A bonus would be given if you finished the both.

1.3 Stochastic Process

We toss a fair coin for a number of times and use H(head) and T(tail) to denote the two sides of the coin. Please compute the expected number of tosses we need to observe a first time occurrence of the following consecutive pattern

$$H, \underbrace{T, T, ..., T}_{k}. \tag{13}$$

Solution: we asume that E is the expection of the consecutive pattern $H, \underbrace{T, T, ..., T}_{k}$, and E_T^k is the expection of $\underbrace{T, T, ..., T}_{k}$. Consider an equivalent form of this pattern $H, \underbrace{T, T, ..., T}_{k-1}, T$, we have

$$\begin{cases} E = 1 + \frac{1}{2}E + \frac{1}{2}E_T^k, \\ E_T^k = E_T^{k-1} + 1 + \frac{1}{2}E_T^k + \frac{1}{2} \times 0. \quad E_T^1 = 2 \end{cases} \tag{14}$$

which $E = 1 + \frac{1}{2}E + \frac{1}{2}E_T^k$ shows the expection of the first toss. At the first time, you may get H or T with the $\frac{1}{2}$ probability. If you got H, OK, you succuced and then you will try to get k times T, the expection will be $\frac{1}{2}E_T^k$; If you got T, you fail and will restart to tosses and the expection will be $\frac{1}{2}E$.
which $E_T^k = E_T^{k-1} + 1 + \frac{1}{2}E_T^k + \frac{1}{2} \times 0$ shows the the expection of the $k-1$ times of T (E_T^{k-1}) and the last toss. At the last toss, as for the first time, you will get H or T with the $\frac{1}{2}$ probability. If you got H, you fail and you need to get k times T over again and the expection will be $\frac{1}{2}E_T^k$. If you got T, OK, you win the game, the expection will be $\frac{1}{2}E_T^k$;
Next, we solve the recursive function above

$$E_T^k = 2^{k+1} - 2 \tag{15}$$

$\Rightarrow$

$$E = 1 + \frac{1}{2}E + \frac{1}{2}(2^{k+1} - 2) \tag{16}$$

$\Rightarrow$

$$E = 2^{k+1} \tag{17}$$

So the expected number of tosses is 2^{k+1}.

Table 1:

ID	Features	Batch Size	Learning Rate	Activation	Regu Rate	Network Shape
1	x1, x2	10	0.03	Tanh	0	4,2

1.4 Probability

Suppose $p \sim Beta(p|\alpha, \beta)$ and $x|p \sim Bernoulli(x|p)$. Show that $p|x \sim Beta(p|\alpha + x, \beta + 1 - x)$, which implies that the Beta distribution can serve as a conjugate prior to the Bernoulli distribution.

Solution: Consider calculating the posterior $p|x$, and we know the likelihood function $x|p$ and the prior p, here we use Bayes' theorem:

$$\begin{aligned} P(p|x) &= \frac{P(x|p)P(p)}{P(x)} \\ &= \frac{P(x|p)P(p)}{\int P(x|p')P(p')dp'} \end{aligned} \tag{18}$$

From the definition, $P(p) \sim Beta(p|\alpha, \beta)$ and $P(x|p) \sim Bernoulli(x|p)$, and the Beta function is

$$Beta(p|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1} \tag{19}$$

so $P(p|x)$ should be

$$\begin{aligned} P(p|x) &= \frac{P(x|p)P(p)}{\int_0^1 P(x|p')P(p')dp'} \\ &= \frac{\binom{m}{n} p^m (1-p)^{n-m} \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1 \binom{m}{n} p^m (1-p)^{n-m} \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} dp} \\ &= \frac{p^{\alpha+m-1}(1-p)^{\beta-1+n-m}}{\int_0^1 p^{\alpha+m-1}(1-p)^{\beta-1+n-m} dp} \\ &= \frac{p^{\alpha+m-1}(1-p)^{\beta-1+n-m}}{B(\alpha+m, \beta+n-m)} \\ &= Beta(p|\alpha+m, \beta+n-m) \end{aligned} \tag{20}$$

So, it implies that the Beta distribution can serve as a conjugate prior to the Bernoulli distribution.

2 Deep Neural Networks: Have a Try

3 Clustering: Mixture of Multinomials

3.1 MLE for multinomial

Derive the maximum-likelihood estimator for the parameter $\mu = (\mu_i)_{i=1}^d$ of a multinomial distribution:

$$P(x|\mu) = \frac{n!}{\prod_i x_i!} \prod_i \mu_i^{x_i}, i = 1, ..., d \tag{21}$$

where $x_i \in \mathbb{N}, \sum_i x_i = n$ and $0 < \mu_i < 1, \sum_i \mu_i = 1$.

Solution: Consider there is a dataset D contains $\mathbb{N}$ documents, the probability is :

$$
\begin{aligned}
P(D|\mu) &= \prod_{j=1}^{N} P(x|\mu) \\
&= \prod_{j=1}^{N} \frac{n!}{\prod_i x_{ji}!} \prod_i \mu_i^{x_{ji}} \\
&= \prod_{j=1}^{N} \frac{n!}{\prod_i x_{ji}!} \bullet \prod_i \mu_i^{\sum_{j=1}^{N} x_{ji}}
\end{aligned} \tag{22}
$$

Maximize the log-likelihood function :

$$
\begin{aligned}
\mathcal{L}(\mu) &= \log P(D|\mu) \\
&= \sum_{j=1}^{N} \log P(x|\mu)
\end{aligned} \tag{23}
$$

and the Lagrange multiplier:

$$
\begin{aligned}
L &= \mathcal{L}(\mu) + \lambda(\sum_{i=1}^{d} \mu_i - 1) \\
&= \sum_{j=1}^{N} \log \frac{n!}{\prod_i x_{ji}!} + \sum_{i=1}^{d} \sum_{j=1}^{N} x_{ij} \log \mu_i + \lambda(\sum_{i=1}^{d} \mu_i - 1)
\end{aligned} \tag{24}
$$

$$
\frac{\partial(\mathcal{L})}{\partial(\mu_i)} = 0 \tag{25}
$$

$\Rightarrow$

$$
\frac{\sum_{j=1}^{N} x_{ij}}{\mu_i} + \lambda = 0 \tag{26}
$$

$\Rightarrow$

$$
\mu_i = -\frac{\sum_{j=1}^{N} x_{ij}}{\lambda} \tag{27}
$$

Because of

$$
\sum_{i=1}^{d} \mu_i = 1
$$

$\Rightarrow$

$$
\sum_{i=1}^{d} -\frac{\sum_{j=1}^{N} x_{ij}}{\lambda} = 1 \tag{28}
$$

$\Rightarrow$

$$
\lambda = -N \tag{29}
$$

$\Rightarrow$

$$
\mu_i = \frac{\sum_{j=1}^{N} x_{ij}}{N} \tag{30}
$$

3.2 EM for mixture of multinomials