CS4487: Home Assignment №2

Department of Computer Science
City University of Hong Kong

Due on October 9, 2019, 7pm

Exercise 1

[5 points]. Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event. Let x has a Poisson distribution,

$$p(x = k|\lambda) = \frac{1}{k!}e^{-\lambda}\lambda^k, \tag{1}$$

where k is an occurrence number and the parameter λ is the average number of events, and the mean and variance are the same $\mathbb{E}[x] = \mathrm{Var}(x) = \lambda$.

a) [2 points]. Derive the maximum-likelihood estimate of λ, given a set of independent and identically distributed (i.i.d.) samples $\mathcal{D} = \{k^{(1)}, \ldots, k^{(m)}\}$.

b) [3 points]. The following table lists the number of intervals (maybe per minute) that are observed to have k occurrences. The total number of intervals is 230. Please calculate the maximum likelihood estimate $\lambda^\star$.

Number of occurrences (k)	0	1	2	3	4 and over
Number of intervals with k	100	81	34	9	6

Exercise 2

[5 points]. Consider the nonlinear error surface $\ell(u, v) = (ue^v - 2ve^{-u})^2$. We start at the point $(u, v) = (1, 1)$ and minimize this error using gradient descent in the u, v space. Use $\alpha = 0.1$ (*i.e.*, learning rate).

ZHANG Deheng
55199998.

CS4487 Assignment 2.

1.

Exercise 1

[5 points]. Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event. Let x has a Poisson distribution,

$$p(x = k|\lambda) = \frac{1}{k!}e^{-\lambda}\lambda^k, \tag{1}$$

where k is an occurrence number and the parameter λ is the average number of events, and the mean and variance are the same $\mathbb{E}[x] = \mathrm{Var}(x) = \lambda$.

a) [2 points]. Derive the maximum-likelihood estimate of λ, given a set of independent and identically distributed (i.i.d.) samples $\mathcal{D} = \{k^{(1)}, \ldots, k^{(m)}\}$.

a)
$$P(D|\lambda) = \prod_{i=1}^{m} P(k^{(i)}|\lambda)$$
$$\propto L(\lambda) = \log \prod_{i=1}^{m} P(k^{(i)}|\lambda)$$
$$L(\lambda) = \sum_{i=1}^{m} \log P(k^{(i)}|\lambda)$$
$$L(\lambda) = \sum_{i=1}^{m} \log\left[\frac{1}{k^{(i)}!} e^{-\lambda} \lambda^{k^{(i)}}\right]$$
$$L(\lambda) = \sum_{i=1}^{m} \left(-\lambda + k^{(i)}\log\lambda - \log(k^{(i)}!)\right)$$

It's fine to maximize P by maximize L, to maximize L, we have.

$$\frac{dL}{d\lambda} = \frac{\sum_{i=1}^{m} k^{(i)}}{\lambda} - \sum_{i=1}^{m}(1) = 0.$$

$$\lambda = \frac{\sum_{i=1}^{m} k^{(i)}}{m}.$$

b) [3 points]. The following table lists the number of intervals (maybe per minute) that are observed to have k occurrences. The total number of intervals is 230. Please calculate the maximum likelihood estimate $\lambda^\star$.

Number of occurrences (k)	0	1	2	3	4 and over
Number of intervals with k	100	81	34	9	6

b)
$$\lambda^\star = \frac{\sum_{i=1}^{m} k^{(i)}}{m} = \frac{\sum_{j=1}^{n} t_j k^{(j)}}{\sum_{j=1}^{n} t_j} = \frac{100\times0 + 81\times1 + 34\times2 + 9\times3 + 6\times4}{230}$$
$$= \frac{200}{230} = \frac{20}{23} \approx 0.8696 = 86.96\%$$

2.

Exercise 2

[5 points]. Consider the nonlinear error surface $\ell(u, v) = (ue^v - 2ve^{-u})^2$. We start at the point $(u, v) = (1, 1)$ and minimize this error using gradient descent in the u, v space. Use $\alpha = 0.1$ (*i.e.*, learning rate).

a) [3 points]. What is the partial derivative of $\ell(u, v)$ with respect to u?

a)
$$\frac{\partial \ell}{\partial u} = 2(ue^v - 2ve^{-u})\left[e^v - (-2ve^{-u})\right]$$
$$= 2(ue^v - 2ve^{-u})(e^v + 2ve^{-u})$$
$$\frac{\partial \ell}{\partial v} = 2(ue^v - 2ve^{-u})(ue^v - 2e^{-u})$$

$$2(ue^{2v} + 2uve^{v-u} - 2ve^{v-u} - 4v^2e^{-2u})$$

b) [1 point]. How many iterations does it take for the error $\ell(u, v)$ to fall below 10^{-14} for the first time? In your programs, make sure to use double precision to get the needed accuracy.

It takes 10 iterations to fall below 10^{-14}.

c) [1 point]. After running enough iterations such that the error has just dropped below 10^{-14}, what is the final (u, v) you get in problem b).

```
13.695429931968397 7.860794463754201
-4.0006368354810276 7.218545990683856
-0.7700242384773525 -3.857237898703024
0.4187828838411616 -1.0704369569657015
0.17803655193163354 -0.3365645910367823
0.02841707522285108 -0.050093524700481326
0.002434440700882361 -0.004250447736045407
0.00018272988668410382 -0.0003187832621601303
1.3552231699862933e-05 -2.364125634816484e-05
1.00419725016074e-06 -1.7517689314309386e-06
Iteration: 10
u: 0.04473629039778207
v: 0.023958714099141746
val: 1.2086833944220747e-15
```

$u \approx 0.044736$
$v \approx 0.023959$

```
1: test.py
    du=2.0*(u*exp(v)-2.0*v*exp(-u))*(exp(v)+2*v*exp(-u))
    return du
def dl_dv(u,v):
    dv=2.0*(u*exp(v)-2.0*v*exp(-u))*(u*exp(v)-2*exp(-u))
    return dv
u = 1.0
v = 1.0
alpha = 0.1
val = l(u,v)
cnt = 0
while(val>=1e-14):
    # 1
    du = dl_du(u,v)
    dv = dl_dv(u,v)
    u -= alpha*du
    v -= alpha*dv
    print(du, dv)
    # 2
#     u -= (alpha*dl_du(u,v))
    #  v -= (alpha*dl_dv(u,v))
    #  print(dl_du(u,v), dl_dv(u,v))

    val = l(u,v)
    cnt += 1

print("Iteration: "+str(cnt)+"\nu: "+str(u)+"\nv: "+str(v)+"\nval: "+str(val))
```

3.

Exercise 3

[5 points]. In Section 3.4 of the lecture note, we have derived the closed-form solution for the ordinary least squares

$$w^{\mathrm{LR}} = \arg\min_w \frac{1}{2}\|Xw - y\|_2^2 = (X^TX)^{-1}X^Ty. \tag{2}$$

With similar arguments, derive the closed-form solution for the ridge regression

$$w^{\mathrm{RR}} = \arg\min_w \frac{1}{2}\|Xw - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2 = (X^TX + \lambda I)^{-1}X^Ty. \tag{3}$$

$$\ell(\vec{w}) = (\vec{y} - X\vec{w})^T(\vec{y} - X\vec{w}) + \lambda\vec{w}^T\vec{w}$$
$$\nabla_{\vec{w}}\ell = \nabla_{\vec{w}}(\vec{y}^T\vec{y} - \vec{y}^TX\vec{w} - (X\vec{w})^T\vec{y} + (X\vec{w})^TX\vec{w})$$
$$+ \lambda\nabla_{\vec{w}}(\vec{w}^T\vec{w})$$
$$= \nabla_{\vec{w}}(\vec{y}^T\vec{y} - 2\vec{y}^TX\vec{w} + \vec{w}^TX^TX\vec{w})$$
$$+ \lambda\nabla_{\vec{w}}(\vec{w}^T\vec{w})$$
$$= \nabla_{\vec{w}}(\vec{w}^TX^TX\vec{w} - 2\vec{y}^TX\vec{w}) + \lambda\nabla_{\vec{w}}(\vec{w}^T\vec{w})$$

$= 2X^TX\vec{w} - 2X^T\vec{y} + 2\lambda I\vec{w}$

For optimal value $\vec{w}^*$

$\nabla_{\vec{w}} L(\vec{w}^*) = 0.$

$X^TX\vec{w}^* - X^T\vec{y} + \lambda I\vec{w}^* = 0.$

$(X^TX + \lambda I)\vec{w} = X^T\vec{y}$

$\vec{w} = (X^TX + \lambda I)^{-1}X^T\vec{y}.$

a) [3 points]. What is the partial derivative of $\ell(u, v)$ with respect to u?

b) [1 point]. How many iterations does it take for the error $\ell(u, v)$ to fall below 10^{-14} for the first time? In your programs, make sure to use double precision to get the needed accuracy.

c) [1 point]. After running enough iterations such that the error has just dropped below 10^{-14}, what is the final (u, v) you get in problem b).

Exercise 3

[5 points]. In Section 3.4 of the lecture note, we have derived the closed-form solution for the ordinary least squares

$$w^{\mathrm{LR}} = \arg\min_{w} \frac{1}{2}\|Xw - y\|_2^2 = (X^T X)^{-1} X^T y. \tag{2}$$

With similar arguments, derive the closed-form solution for the ridge regression

$$w^{\mathrm{RR}} = \arg\min_{w} \frac{1}{2}\|Xw - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2 = (X^T X + \lambda I)^{-1} X^T y. \tag{3}$$