Home Assignment №2 Solutions

November 5, 2024

Exercise 1

[3 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) [1 point]. Derive the maximum-likelihood estimate of λ, given a set of independent and identically distributed (i.i.d.) samples $\mathcal{D} = \{k^{(1)}, \dots, k^{(M)}\}$.

b) [2 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

Solution:

a) The log-likelihood of the data $\mathcal{D} = \{k^{(1)}, \dots, k^{(M)}\}$ is

$$L(\lambda) = \log p(\mathcal{D}; \lambda) = \sum_{i=1}^{M} \log p(x^{(i)}; \lambda) = \sum_{i=1}^{M} \log \frac{1}{k^{(i)}!} e^{-\lambda} \lambda^{k^{(i)}} \tag{2}$$

$$= \sum_{i=1}^{M} [-\lambda + k^{(i)} \log \lambda - \log k^{(i)}!] \tag{3}$$

$$= -M\lambda + \left(\sum_{i=1}^{M} k^{(i)} \right) \log \lambda - \sum_{i=1}^{M} \log k^{(i)}! \,. \tag{4}$$

We take the derivative of $L(\lambda)$ with respect to λ and set it to zero,

$$\frac{\partial L(\lambda)}{\partial \lambda} = -M + \frac{1}{\lambda} \sum_{i=1}^{M} k^{(i)} = 0, \tag{5}$$

$$\lambda^\star = \frac{1}{M} \sum_{i=1}^{M} k^{(i)}, \tag{6}$$

b) The total number of intervals is 230. Assuming that the observation for the "4 and over" was actually 4, the maximum likelihood estimate for the Poisson distribution is

$$\lambda^\star = \frac{1}{230}(100(0) + 81(1) + 34(2) + 9(3) + 6(4)) = 0.8696. \tag{7}$$

Exercise 2

[3 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) [1 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)? Round your answer to the thousandths place.

Solution:

a)

$$\frac{\partial \ell(u,v)}{\partial u} = 2(ue^{v} - 2ve^{-u})(e^{v} + 2ve^{-u}). \tag{8}$$

b) 10 iterations.

c) $(0.045, 0.024)$.

```
# Reference codes for Exercise 2 problem b) and c)
from random import random
import math

def gradient_decent(fn, partial_derivatives,
n_variables, lr=0.1,max_iter=20, tolerance=1e-14):
    theta = [1, 1]
    y_cur = fn(*theta)
    for i in range(max_iter):
        # Calculate gradient of current theta.
        gradient = [f(*theta) for f in partial_derivatives]
        # Update the theta by the gradient.
        for j in range(n_variables):
            theta[j] -= gradient[j] * lr
        # Check if converged or not.
        y_cur, y_pre = fn(*theta), y_cur
        if (y_cur < tolerance):
            print(i+1) # i starts from 0
            break
    return theta

def f(u, v):
    return ((u*math.exp(v) - 2*v*math.exp(-u))**2)

def df_du(u, v):
    return 2 * (u*math.exp(v) - 2*v*math.exp(-u))
    * (math.exp(v) + 2 * v * math.exp(-u))

def df_dv(u, v):
    return 2 * (u*math.exp(v) - 2*v*math.exp(-u))
    * (u*math.exp(v) - 2 * math.exp(-u))

def main():
    print("Solve the minimum value of error surface function:")
    n_variables = 2
    para = gradient_decent(f, [df_du, df_dv],
    n_variables)
    para = [round(x, 3) for x in para]
    print("The solution is: (u, v): %s\n" % (para))
```

Exercise 3

[3 points]. Suppose that $x \in \mathbb{R}^2 \left(x = [x_1, x_2]^T\right)$. Consider the following optimization problem:

$$\begin{aligned} \min_{x} \quad & x_1^2 + x_2^2 \\ \text{s.t.} \quad & (x_1 - 1)^2 + (x_2 - 1)^2 \leq 1 \\ & (x_1 - 1)^2 + (x_2 + 1)^2 \leq 1. \end{aligned}$$

a) [1 point]. Sketch the feasible set and level sets of the objective. Find the optimal point $x^\star$ and optimal value $p^\star$.

b) [1 point]. Give the KKT conditions. Do there exist Lagrange multipliers $\lambda_1^\star$ and $\lambda_2^\star$ that prove that $x^\star$ is optimal?

c) [1 point]. Derive and solve the Lagrange dual problem. Does strong duality hold?

Solution:

a) The feasible set is plotted below. The figure shows the feasible set (the intersection of the two shaded disks) and some contour lines of the objective function. There is only one feasible point, $x^\star = [1, 0]^T$, so it is optimal for the primal problem, and we have $p^\star = 1$.

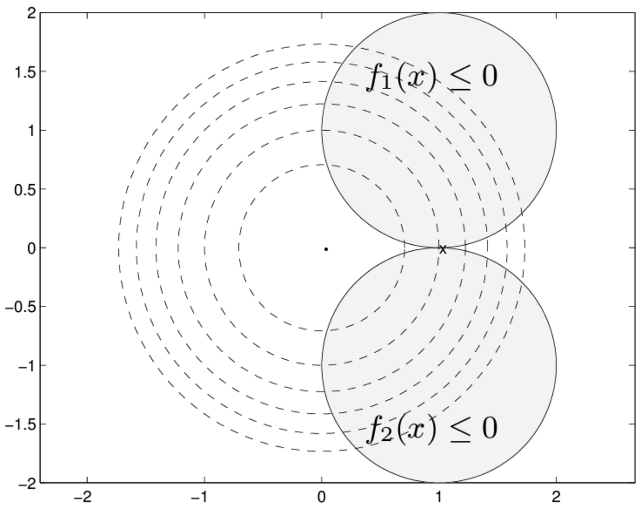

Figure 1: Feasible set

b) Let's write down Lagrangian with λ_1 and λ_2 as Lagrangian multipliers:

$L(x_1, x_2, \lambda_1, \lambda_2) = x_1^2 + x_2^2 + \lambda_1((x_1-1)^2 + (x_2-1)^2 - 1) + \lambda_2((x_1-1)^2 + (x_2+1)^2 - 1).$

Hence, the KKT conditions are:
Primal feasibility:

$$(x_1-1)^2 + (x_2-1)^2 - 1 \le 0,$$
$$(x_1-1)^2 + (x_2+1)^2 - 1 \le 0.$$

Dual feasibility:

$$\lambda_1 \ge 0,$$
$$\lambda_2 \ge 0.$$

Complementary slackness:

$$\lambda_1((x_1-1)^2 + (x_2-1)^2 - 1) = 0,$$
$$\lambda_2((x_1-1)^2 + (x_2+1)^2 - 1) = 0.$$

Lagrangian stationary:

$$\frac{\partial L}{\partial x_1} = 2x_1 + 2\lambda_1(x_1-1) + 2\lambda_2(x_1-1) = 0,$$
$$\frac{\partial L}{\partial x_2} = 2x_2 + 2\lambda_1(x_2-1) + 2\lambda_2(x_2+1) = 0.$$

At $x = [1,0]^T$, these conditions reduce to

$$\begin{aligned} \lambda_1 &\geq 0, \\ \lambda_2 &\geq 0, \\ 2 &= 0, \\ -2\lambda_1 + 2\lambda_2 &= 0, \end{aligned}$$

which (clearly, in view of the third equation) have no solution.

c) The Lagrange dual function is given by

$$g(\lambda_1, \lambda_2) = \inf_{x_1, x_2} L(x_1, x_2, \lambda_1, \lambda_2),$$

where

$$\begin{aligned} &L(x_1, x_2, \lambda_1, \lambda_2) \\ &\quad = x_1^2 + x_2^2 + \lambda_1\left((x_1-1)^2 + (x_2-1)^2 - 1\right) + \lambda_2\left((x_1-1)^2 + (x_2+1)^2 - 1\right) \\ &\quad = (1+\lambda_1+\lambda_2)\, x_1^2 + (1+\lambda_1+\lambda_2)\, x_2^2 - 2(\lambda_1+\lambda_2)\, x_1 - 2(\lambda_1-\lambda_2)\, x_2 + \lambda_1 + \lambda_2. \end{aligned}$$

L reaches its minimum for

$$x_1 = \frac{\lambda_1+\lambda_2}{1+\lambda_1+\lambda_2}, \quad x_2 = \frac{\lambda_1-\lambda_2}{1+\lambda_1+\lambda_2}$$

and we find

$$g(\lambda_1, \lambda_2) = \begin{cases} -\frac{(\lambda_1+\lambda_2)^2 + (\lambda_1-\lambda_2)^2}{1+\lambda_1+\lambda_2} + \lambda_1 + \lambda_2 & , 1+\lambda_1+\lambda_2 \geq 0 \\ -\infty & \text{, otherwise} \end{cases}$$

where we interpret $a/0 = 0$ if $a = 0$ and as $-\infty$ if $a < 0$. The Lagrange dual problem is given by

$$\begin{aligned} \max_{\lambda_1, \lambda_2} \quad & \left(\lambda_1 + \lambda_2 - (\lambda_1 - \lambda_2)^2\right) / (1 + \lambda_1 + \lambda_2) \\ \text{s.t.} \quad & \lambda_1, \lambda_2 \geq 0 \end{aligned}$$

Since g is symmetric, the optimum (if it exists) occurs with $\lambda_1 = \lambda_2$. The dual function then simplifies to

$$g\left(\lambda_1, \lambda_1\right) = \frac{2\lambda_1}{2\lambda_1 + 1}$$

We see that $g\left(\lambda_1, \lambda_2\right)$ tends to 1 as $\lambda_1 \to \infty$. We have $d^\star = p^\star = 1$, but the dual optimum is not attained. Recall that the KKT conditions only hold if (1) strong duality holds, (2) the primal optimum is attained, and (3) the dual optimum is attained. In this example, the KKT conditions fail because the dual optimum is not attained.

Exercise 4

[3 points]. In the formulation of SVM, we need to compute the margin (i.e., the distance) between an arbitrary point $x^{(i)}$ in the N-dimensional space and a hyperplane $w^T x + b = 0$, which can be formulated as the following optimization problem:

$$\begin{aligned} \min_{x} \quad & \left\| x^{(i)} - x \right\|_2 \\ \text{s.t.} \quad & w^T x + b = 0. \end{aligned}$$

a) [1 point]. Is this problem convex and why?

b) [2 points]. Using the Lagrange duality to solve for the optimal x and the distance. (Remember to form the Lagrangian and derive the Lagrange dual function).

Solution:

a) Yes, because the objective function, L2 norm of an affine function, is convex, and the equation constraint is affine.

b) The given convex optimization problem is equivalent to

$$\begin{aligned} \min_{x} \quad & \frac{1}{2} \left\| x^{(i)} - x \right\|_2^2 \\ \text{s.t.} \quad & w^T x + b = 0. \end{aligned}$$

We form the Lagrangian:

$$\begin{aligned} L(x,\lambda) &= \frac{1}{2}\left(x - x^{(i)}\right)^T \left(x - x^{(i)}\right) + \lambda\left(w^T x + b\right) \\ &= \frac{1}{2}x^T x + \left(\lambda w - x^{(i)}\right)^T x + \frac{1}{2}\left(x^{(i)}\right)^T x^{(i)} + \lambda b. \end{aligned} \tag{9}$$

Taking the derivative of L w.r.t. x and setting it to zero, we have

$$\begin{aligned} \frac{\partial L(x,\lambda)}{\partial x} &= x + \lambda w - x^{(i)} = 0 \\ x &= x^{(i)} - \lambda w. \end{aligned} \tag{10}$$

Plugging Eq. (10) back to the Lagrangian in Eq. (9), we arrive at the Lagrange dual function:

$$\begin{aligned} g(\lambda) &= \frac{1}{2}w^T w \lambda^2 + \lambda\left(w^T\left(x^{(i)} - \lambda w\right) + b\right) \\ &= -\frac{1}{2}w^T w \lambda^2 + \left(w^T x^{(i)} + b\right)\lambda. \end{aligned}$$

By maximizing $g(\lambda)$, we are able to solve for $\lambda^\star$:

$$\begin{aligned} \frac{\partial g(\lambda)}{\partial \lambda} &= -w^T w \lambda^\star + \left(w^T x^{(i)} + b\right) = 0 \\ \lambda^\star &= \frac{w^T x^{(i)} + b}{w^T w} \end{aligned}$$

Finally, we are able to compute the distance:

$$\left\|x - x^{(i)}\right\|_2 = \|\lambda^\star w\|_2 = |\lambda^\star| \cdot \|w\|_2 = \frac{\left|w^T x^{(i)} + b\right|}{\|w\|_2^2}\|w\|_2 = \frac{\left|w^T x^{(i)} + b\right|}{\|w\|_2},$$

as desired.

Exercise 5

[3 points]. In the lecture note, we have given a detailed derivation of the dual form of SVM with soft margin. With simpler arguments, derive the dual form of SVM with hard margin

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}w^T w \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, M. \end{aligned}$$

Compare the two dual forms.

Solution: We form the Lagrangian of the above problem

$$\begin{aligned} L(w,b,\alpha) &= \frac{1}{2}w^T w + \sum_{i=1}^{M} \alpha_i (1 - y^{(i)}(w^T x^{(i)} + b)) \\ &= \frac{1}{2}w^T w - \sum_{i=1}^{M} \alpha_i y^{(i)} w^T x^{(i)} - \sum_{i=1}^{M} \alpha_i y^{(i)} b + \sum_{i=1}^{M} \alpha_i . \end{aligned}$$

Then, we take the partial derivatives of L w.r.t. w, b and set them to zero

$$\nabla_w L = \nabla_w \left(\frac{1}{2}w^T w - \sum_{i=1}^{M} \alpha_i y^{(i)} w^T x^{(i)} \right) = w - \sum_{i=1}^{M} \alpha_i y^{(i)} x^{(i)} = 0, \tag{11}$$

$$\nabla_b L = \nabla_b \left(- \sum_{i=1}^{M} \alpha_i y^{(i)} b \right) = - \sum_{i=1}^{M} \alpha_i y^{(i)} = 0. \tag{12}$$

Plugging Eq. (11) and Eq. (12) back into the Lagrangian, we obtain

$$\begin{aligned} g(\alpha) = \frac{1}{2} \left(\sum_{i=1}^{M} \alpha_i y^{(i)} x^{(i)} \right)^T \left(\sum_{j=1}^{M} \alpha_j y^{(j)} x^{(j)} \right) - \sum_{i=1}^{M} \alpha_i y^{(i)} \left(\sum_{j=1}^{M} \alpha_j y^{(j)} x^{(j)} \right)^T x^{(i)} + \sum_{i=1}^{M} \alpha_i \\ = \sum_{i=1}^{M} \alpha_i + \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)} - \sum_{i=1}^{M} \sum_{j=1}^{M} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)} \\ = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)}. \qquad (13) \end{aligned}$$

Putting $g(\alpha)$ together with the constraints, we obtain the following dual optimization problem:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)} \\ \text{s.t.} \quad & \sum_{i=1}^{M} \alpha_i y^{(i)} = 0, \\ & \alpha_i \geq 0, \quad i = 1, \ldots, M. \end{aligned} \tag{14}$$

Recall that the dual problem of SVM with soft margin is

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)} \\ \text{s.t.} \quad & \sum_{i=1}^{M} \alpha_i y^{(i)} = 0, \\ & 0 \le \alpha_i \le C, \quad i = 1, \ldots, M. \end{aligned} \tag{15}$$

Apparently the ranges of α_i are different.