Statistical Machine Learning: Homework #1

Professor Devika Subramanian

Marco Lagos

Problem 1

See **sampler.ipynb**

Problem 2

Prove that the sum of two independent Poisson random variables is also a Poisson random variable.

Solution
Consider two independent random variables s.t. $X \sim Poi(\lambda_1)$ and $Y \sim Poi(\lambda_2)$. Let $Z = X + Y$.
Let $\Omega_X = \Omega_Y = \Omega_Z = \{1, 2, ...\}$. The convolution formula for discrete distributions is (for $n \in \Omega_Z$ and $i \leq n$):

$$p_Z(n) = \sum_{i=0}^{n} p_X(i) p_Y(n-i)$$

In addition, the binomial theorem is as follows:

$$(a+b)^n = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} a^{n-k} b^k$$

From the following, we can conclude that $Z \sim Poi(\lambda_1 + \lambda_2)$:

$$\begin{aligned}
p_Z(n) &= \sum_{i=0}^{n} p_X(i) p_Y(n-i) \\
&= \sum_{i=0}^{n} e^{-\lambda_1} \frac{\lambda_1^i}{i!} \cdot e^{-\lambda_2} \frac{\lambda_2^{n-i}}{(n-i)!} \\
&= e^{-(\lambda_1+\lambda_2)} \sum_{i=0}^{n} \frac{\lambda_1^i \lambda_2^{n-i}}{i!(n-i)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{i=0}^{n} \frac{n!}{i!(n-i)!} \lambda_1^i \lambda_2^{n-i} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} (\lambda_1 + \lambda_2)^n
\end{aligned}$$

Problem 3

Let A, B, C be events. Show that if $P(A|B,C) > P(A|B)$ then $P(A|B,C^c) < P(A|B)$. Here C^c denotes the complement of C. Assume that each event we are conditioning on has positive probability.

Solution
By the conditional probability formula:

$$P(A|B,C) = \frac{P(A \cap B \cap C)}{P(B \cap C)}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

And by the complement rule:

$$P(C^c) = 1 - P(C)$$

From the following, we can conclude that $P(A|B,C) > P(A|B) \rightarrow P(A|B,C^c) < P(A|B)$:

$$\begin{aligned} P(A|B,C) &> P(A|B) \\ \frac{P(A \cap B \cap C)}{P(B \cap C)} &> \frac{P(A \cap B)}{P(B)} \\ P(A \cap B \cap C) &> P(A \cap B) \end{aligned}$$

Substituting $P(A \cap B \cap C^c) = P(A \cap B) - P(A \cap B \cap C)$:

$$\begin{aligned} P(A \cap B \cap C^c) &< P(A \cap B) \\ \frac{P(A \cap B \cap C^c)}{P(B)} &< \frac{P(A \cap B)}{P(B)} \\ \frac{P(A \cap B \cap C^c)}{P(B)} &< P(A|B) \\ P(A|B,C^c) &< P(A|B) \end{aligned}$$

Problem 4

Consider the vectors $u = \begin{bmatrix} 1 & 2 \end{bmatrix}^T$ and $v = \begin{bmatrix} 2 & 3 \end{bmatrix}^T$. Define the matrix $M = uv^T$. Compute the eigenvalues and eigenvectors of M.

Solution
Computing M:

$$\begin{aligned} M &= uv^T \\ &= \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 2 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 2 & 3 \\ 4 & 6 \end{bmatrix} \end{aligned}$$

Problem 4 continued on next page...

Getting the eigenvalues from the characteristic equation:

$$\begin{aligned} det(A - \lambda I) &= 0 \\ det \begin{bmatrix} 2-\lambda & 3 \\ 4 & 6-\lambda \end{bmatrix} &= 0 \\ (2-\lambda)(6-\lambda) - 3 \cdot 4 &= 0 \\ 12 - 8\lambda + \lambda^2 - 12 &= 0 \\ \lambda^2 - 8\lambda &= 0 \\ \lambda(\lambda - 8) &= 0 \\ \lambda_1 = 8, \lambda_2 &= 0 \end{aligned}$$

The eigenvalues are then $\lambda_1 = 8$ and $\lambda_2 = 0$. Now, by solving the equation $(A - \lambda I)v = 0$ for each eigenvalue, we find the corresponding eigenvectors:

$$v_{\lambda_1} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$v_{\lambda_2} = \begin{bmatrix} -3 \\ 2 \end{bmatrix}$$

Problem 5

Let $A \in \mathbb{R}^{n \times n}$ be symmetric matrix. We say that A is positive semi-definite if $\forall x \in \mathbb{R}^n, x^T Ax \geq 0$. Show that if A is positive semi-definite, then all eigenvalues of A are non-negative.

Solution

Consider an eigenvector v of A so that $Av = \lambda v$:

$$\begin{aligned} Av &= \lambda v \\ v^T Av &= \lambda v^T v \end{aligned}$$

Since A is positive semi-definite, we know that $x^T Ax \geq 0$. In addition, $v^T v$ is the squared norm of vector v so $v^T v \geq 0$.

$$\begin{aligned} x^T Ax &\geq 0 \\ \lambda v^T v &\geq 0 \end{aligned}$$

If λ were negative then $\lambda v^T v$ would also be negative since $v^T v$ is non-negative, but we have already established that $v^T v \geq 0$. Therefore, we can conclude that all eigenvalues of A are non-negative.

Problem 6

Provide one example for each of the following cases, where A, B are 2×2 matrices.

1. $(A + B)^2 \neq A^2 + 2AB + B^2$
2. $AB = 0, A \neq 0, B \neq 0$

Problem 6 continued on next page...

Solution

1. For $(A+B)^2 \neq A^2 + 2AB + B^2$, the left side of the equation is $(A+B)^2$:

$$\begin{aligned}(A+B)^2 &= (A+B)(A+B) \\ &= A(A+B) + B(A+B) \\ &= A^2 + AB + BA + B^2\end{aligned}$$

We must find a matrix such that $AB \neq BA$. It is clear with these two matrices:

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$B = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}$$

2. For $AB = 0, A \neq 0, B \neq 0$, we simply need to find a matrix that cancels each other out during matrix multiplication:

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$B = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}$$

Problem 7

Let u denote a real vector normalized to unit length. That is, $u^T u = 1$. Show that $A = I - 2uu^T$ is orthogonal, i.e., $A^T A = I$.

Solution

$$\begin{aligned}A^T A &= (I - 2uu^T)^T (I - 2uu^T) \\ &= (I^T - (2uu^T)^T)(I - 2uu^T) \\ &= (I - 2uu^T)(I - 2uu^T) \\ &= I^2 - 2Iuu^T - 2uu^T I + 4(uu^T)^2 \\ &= I^2 - 4uu^T + 4(uu^T)^2 \\ &= I^2 - 4 + 4 \\ &= I\end{aligned}$$

Problem 8

A function f is convex on a given set S iff for $\lambda \in [0,1]$ and for all $x, y \in S$, the following holds:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

Moreover, a univariate function $f(x)$ is convex on a set S iff its second derivative $f''(x)$ is non-negative everywhere in the set. Prove the following assertions:

1. $f(x) = x^3$ is convex for $x \geq 0$

Problem 8 continued on next page...

2. $f(x_1, x_2) = max(x_1, x_2)$ is convex on $\mathbb{R}$
3. If univeriate functions f and g are convex on S, then $f + g$ is convex on S
4. If univariate functions f and g are convex and non-negative on S, and have their minimum within S at the same point, then fg is convex on S

Solution
1. $f(x)$ is a univariate function. Let us take the second derivative:

$$f(x) = x^3$$
$$f'(x) = 2x^2$$
$$f''(x) = 6x$$

Given that $6x \geq 0$ for $x \geq 0$, then we can conclude that $f(x) = x^3$ is convex for $x \geq 0$.
2.
3. Given that f, g are univariate functions convex on S, then we also know that $f'', g'' \geq 0$ for all $x \in S$. Consider $h(x) = f(x) + g(x)$ for all $x \in S$. It follows that:

$$h''(x) = f''(x) + g''(x)$$

Since both $f'', g'' \geq 0$ by definition, then $h'' \geq 0$ or in words h'' is non-negative $\forall x \in S$. We can then conclude that $f + g$ is convex on set S.
4.

Problem 9

The entropy of a categorical distribution on K values is defined as

$$H(p) = -\sum_{i=1}^{K} p_i log(p_i)$$

Using the method of Lagrange multipliers, find the categorical distribution that has the highest entropy.

Solution

A categorical distribution is a discrete probability distribution over a finite set of K distinct categories or values. We want to find the categorical distribution that maximizes its entropy. Entropy measures the uncertainty or disorder in a probability distribution.

The method of Lagrange Multipliers is used to find the maximum or minimum of a function subject to some constraints. Our constraint in this case is:

$$\sum_{i=1}^{K} p_i = 1$$

λ (lambda) is the Lagrange multiplier associated with the constraint. Our goal is to maximize L with respect to the probabilities $p_1, p_2, ..., p_K$ and λ. To find the maximum, we set the partial derivatives of L with respect to each variable to zero (for each p_i and λ to enforce the constraint):

$$L(p, \lambda) = -\sum_{i=1}^{K} p_i \log(p_i) + \lambda \left(\sum_{i=1}^{K} p_i - 1 \right)$$

$$\frac{\partial L}{\partial p_i} = -\log(p_i) - 1 + \lambda = 0$$

$$\frac{\partial L}{\partial \lambda} = \sum_{i=1}^{K} p_i - 1 = 0$$

From the λ equation:

$$\sum_{i=1}^{K} p_i = 1$$

From the p_i equation:

$$p_i = e^{-(\lambda+1)}$$

Since the probabilities must sum to 1, we have:

$$K \cdot e^{-(\lambda+1)} = 1$$

Solving for λ:

$$\lambda = -1 - \log\left(\frac{1}{K}\right) = \log(K) - 1$$

Substituting for λ:

$$p_i = e^{-(\log(K)-1)} = \frac{1}{K}$$

We can conclude that the categorical distribution that maximizes entropy is uniform distribution. This makes sense, since all the categories are equally likely:

$$p_i = \frac{1}{K}$$

Problem 10

Consider a linear regression problem in which we want to weight different training examples differently. Specifically, suppose we want to minimize:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} w^{(i)} (\theta^T x^{(i)} - y^{(i)})^2$$

Part A
Show that $J(\theta)$ can be written in the form:

$$J(\theta) = (X\theta - y)^T W (X\theta - y)$$

for an appropriate diagonal matrix W, where X is the $m \times d$ input matrix and y is a $m \times 1$ vector denoting the associated outputs. State clearly what W is.

Solution

We can express $X\theta - y$:

$$X\theta - y = \begin{bmatrix} \theta^T x^{(1)} - y^{(1)} \\ \theta^T x^{(2)} - y^{(2)} \\ \vdots \\ \theta^T x^{(m)} - y^{(m)} \end{bmatrix} = \begin{bmatrix} \theta^T x^{(1)} \\ \theta^T x^{(2)} \\ \vdots \\ \theta^T x^{(m)} \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

We can rewrite $J(\theta)$ as follows:

$$J(\theta) = \frac{1}{2} (X\theta - y)^T \operatorname{diag}(w) (X\theta - y)$$

Here, $\operatorname{diag}(w)$ represents a diagonal matrix with the weights $w^{(i)}$ on the diagonal:

$$\operatorname{diag}(w) = \begin{bmatrix} w^{(1)} & 0 & \cdots & 0 \\ 0 & w^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w^{(m)} \end{bmatrix}$$

Now, let's perform the matrix multiplications and transpose:

$$\begin{aligned} J(\theta) &= \frac{1}{2} (X\theta - y)^T \operatorname{diag}(w) (X\theta - y) \\ &= \frac{1}{2} \left(\begin{bmatrix} \theta^T x^{(1)} - y^{(1)} \\ \theta^T x^{(2)} - y^{(2)} \\ \vdots \\ \theta^T x^{(m)} - y^{(m)} \end{bmatrix} \right)^T \begin{bmatrix} w^{(1)} & 0 & \cdots & 0 \\ 0 & w^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w^{(m)} \end{bmatrix} \begin{bmatrix} \theta^T x^{(1)} - y^{(1)} \\ \theta^T x^{(2)} - y^{(2)} \\ \vdots \\ \theta^T x^{(m)} - y^{(m)} \end{bmatrix} \end{aligned}$$

Now, we have the expression in the desired form, where W is a diagonal matrix:

$$J(\theta) = (X\theta - y)^T \operatorname{diag}(w) (X\theta - y)$$

So, W is a diagonal matrix where the diagonal elements are the weights $w^{(i)}$:

$$W = \operatorname{diag}(w) = \begin{bmatrix} w^{(1)} & 0 & \cdots & 0 \\ 0 & w^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w^{(m)} \end{bmatrix}$$

Part B

If all the $w^{(i)}$'s are equal to 1, the normal equation to solve for the parameter θ is:

$$X^T X\theta = X^T y$$

and the values of θ that minimizes $J(\theta)$ is $(X^T X)^{-1} X^T y$. By computing the derivative of the weighted $J(\theta)$ and setting it equal to zero, generalize the normal equation to the weighted setting and solve for θ in closed form in terms of W, X, and y.

Solution

To generalize the normal equation to the weighted setting and solve for θ in closed form in terms of W, X, and y, we will compute the derivative of $J(\theta)$ with respect to θ and set it equal to zero:

$$\begin{aligned}\frac{\partial J(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta}\left(\frac{1}{2}\sum_{i=1}^{m} w^{(i)}(\theta^T x^{(i)} - y^{(i)})^2\right) \\ &= \frac{1}{2}\sum_{i=1}^{m} 2w^{(i)}(\theta^T x^{(i)} - y^{(i)})\frac{\partial}{\partial \theta}(\theta^T x^{(i)} - y^{(i)}) \\ &= \sum_{i=1}^{m} w^{(i)}(\theta^T x^{(i)} - y^{(i)})x^{(i)}\end{aligned}$$

Setting the derivative equal to zero and solve for θ:

$$\begin{aligned}0 &= \sum_{i=1}^{m} w^{(i)}(\theta^T x^{(i)} - y^{(i)})x^{(i)} \\ 0 &= \sum_{i=1}^{m} w^{(i)}(\theta^T x^{(i)} x^{(i)} - y^{(i)}x^{(i)}) \\ 0 &= \sum_{i=1}^{m} w^{(i)}\theta^T x^{(i)}x^{(i)} - \sum_{i=1}^{m} w^{(i)}y^{(i)}x^{(i)} \\ 0 &= X^T W X\theta - X^T W y \\ \theta &= (X^T W X)^{-1}X^T W y\end{aligned}$$

Part C

To predict the target value for an input vector x, one choice for the weighting functions $w^{(i)}$ is:

$$w^{(i)} = exp\left(-\frac{(x - x^{(i)})^T(x - x^{(i)})}{2\tau^2}\right)$$

Points near x are weighted more heavily than points far away from x. The parameter τ is a bandwidth defining the sphere of influence around x. Note how the weights are defined by the input x. Write down an algorithm for calculating θ by gradient descent for locally weighted linear regression. Is locally weighted linear regression a parametric or a non-parametric method?

Solution

Locally weighted linear regression is a non-parametric method since it does not assume a set relationship between the input and the target variable. It adapts the model based on the structure of the data, weighting data points that are closer more heavily.

To use locallly weighted lienar regression with gradient descent, coefficient vector θ needs to be adjusted based on minimizing the cost function $J(\theta)$:

1. Initialize θ to random values
2. Choose learning rate α and a convergence criteria
3. Repeat the following until convergence:

```
1: function GRADIENT-DESCENT-ON-LWLR(initial, α, iterations)
2:     Initialize θ to initial                          ▷ Initialize model parameters
3:     m ← number of training examples
4:     repeat
5:         for i from 1 to m do                         ▷ For each training example
6:             w^(i) = exp(−(x−x^(i))^T(x−x^(i)) / 2τ^2)
7:             δ^(i) = w^(i)(θ^T x^(i) − y^(i))
8:             θ_j = θ_j − (α/m) Σ_{i=1}^{m} δ^(i) x_j^(i)
```

```
        end for
    until convergence or iterations reached
    return θ                                   ▷ Final model parameters
end function
```

Problem 11

An estimator of an unknown parameter is called unbiased if its expected value equals the true value of the parameter. Here, you will prove that the least-squares estimate given by the normal equation for linear regression is an unbiased estimate of the true parameter θ^*. We first assume that the data:

$$D = \{x^{(i)}, y^{(i)} | 1 \geq i \geq m; x^{(i)} \in \mathbb{R}^d; y^{(i)} \in \mathbb{R}\}$$

comes from the linear model:

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

where each $e^{(i)}$ is an independent random variable drawn from a normal distribution with zero mean and variance σ^2 . When considering the bias of an estimator, we treat the input $x^{(i)}$'s as fixed but arbitrary, and the true parameter vector θ^* as fixed but unknown. Expectations are taken over possible realizations of the output values $y^{(i)}$'s.

Part A
Show that $E[\theta] = \theta^*$ for the least squares estimator.

Solution
The goal is to show that $E[\hat{\theta}] = \theta^*$. This is $\hat{\theta}$:

$$\begin{aligned}
\hat{\theta} &= (X^T X)^{-1} X^T Y \\
E[\hat{\theta}] &= E[(X^T X)^{-1} X^T Y] \\
E[\hat{\theta}] &= (X^T X)^{-1} X^T E[Y]
\end{aligned}$$

Since X and Y are fixed (but random), we can move them outside the expectation. Since $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$:

$$E[Y] = E[\theta^T X + \epsilon] = \theta^T X + E[\epsilon] = \theta^T X$$

Substituting for $E[Y]$:

$$\begin{aligned}
E[\hat{\theta}] &= (X^T X)^{-1} X^T \theta^T X \\
E[\hat{\theta}] &= (X^T X)^{-1} X^T X \theta \\
E[\hat{\theta}] &= \theta
\end{aligned}$$

Part B
Show that the variance of the least squares estimator is $Var(\theta) = (X^T X)^{-1} \sigma^2$.

Solution
The goal is to show $Var(\theta) = (X^T X)^{-1} \sigma^2$. This is $\hat{\theta}$:

$$\begin{aligned}
\hat{\theta} &= (X^T X)^{-1} X^T Y \\
Var[\hat{\theta}] &= Var[(X^T X)^{-1} X^T Y] \\
Var[\hat{\theta}] &= (X^T X)^{-1} X^T Var[Y]((X^T X)^{-1} X^T)^T \\
Var[\hat{\theta}] &= (X^T X)^{-1} X^T Var[Y] X (X^T X)^{-1}
\end{aligned}$$

Since X and Y are fixed (but random), we can move them outside the expectation. Since $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$:

$$Var(Y) = Var(\theta^T X + \epsilon) = Var(\epsilon) = \sigma^2$$

Substituting for $E[Y]$:

$$Var[\hat{\theta}] = (X^T X)^{-1} X^T \sigma^2 X (X^T X)^{-1}$$
$$Var[\hat{\theta}] = (X^T X)^{-1} \sigma^2$$

Problem 12

See **ex1.ipynb**