Solution Set #1

1. (Independent random variables) *Find an example of three binary random variables X, Y and Z such that they are pairwise independent but not mutually independent, i.e., $p(x,y) = p(x)p(y)$, $p(x,z) = p(x)p(z)$, $p(y,z) = p(y)p(z)$ for all $(x,y,z) \in \{0,1\}^3$, but $p(x,y,z) \neq p(x)p(y)p(z)$ for some $(x,y,z) \in \{0,1\}^3$.*

Solution

Consider three binary random variables X, Y, and Z whose joint probability is given by

$$\Pr(X=x, Y=y, Z=z) = \begin{cases} 0.25 & \text{if } x+y+z \equiv 0 \pmod 2 \\ 0 & \text{otherwise} \end{cases}$$

Then, by marginalization, we can check

$$\begin{aligned} p(x) = p(y) = p(z) = 1/2 \quad &\text{for all} \quad (x,y,z) \in \{0,1\}^3 \\ p(x,y) = p(x,z) = p(y,z) = 0.25 \quad &\text{for all} \quad (x,y,z) \in \{0,1\}^3 \end{aligned}$$

Therefore, X, Y and Z are pairwise independent.

However, we can easily check

$$\Pr(X=0, Y=0, Z=0) = 0.25 \neq 0.125 = \Pr(X=0)\Pr(Y=0)\Pr(Z=0)$$

Therefore, X, Y and Z are not mutually independent.

2. (Saddle points)

(a) *Construct a function $f : \mathbb{R}^2 \to \mathbb{R}$ such that its saddle points are given by $\{(n, 2m+1-n) | n \in \mathbb{Z}, m \in \mathbb{Z}\}$, where $\mathbb{Z}$ is the set of integers.*

Solution

There are many solutions satisfying this condition. Here's one example:

$$f(x,y) = \cos(\pi x) + \cos(\pi y).$$

For intergers m and n, if m and n are even, then $f(m,n)$ is a local maximum, and if m and n are odd, $f(m,n)$ is a local minimum. Otherwise, (m,n) is a saddle point of f. Therefore, the saddle points of this function is given by $\{(n, 2m+1-n) | n \in \mathbb{Z}, m \in \mathbb{Z}\}$.

(b) *Construct a function $f : \mathbb{R}^n \to \mathbb{R}$ such that it has roughly 2^n times more saddle points than local minima in $[0, 100]^n$, where $[a,b]$ is the closed interval between a and b.*

Solution

There are many solutions satisfying this condition. Here are two examples:

(1)

$$f(x_1, ..., x_n) = \sum_{i=1}^{n} \cos(\pi x_i)$$

In a similar way to (a), the number of saddle points of f is roughly 100^n and the number of local minima is 50^n. Therefore, there are roughly 2^n times more saddle points than local minima in $[0, 100]^n$.

(2)

$$f(x_1, ..., x_n) = \sum_{i=1}^{n} x_i(x_i - 1)(x_i - 2)$$

Then, the partial derivatives of the function f is,

$$\frac{\partial f}{\partial x_i} = 3{x_i}^2 - 6x_i + 2.$$

Therefore, the critical points of this function are given by $\{(x_1, ..., x_n)|x_i = \frac{3\pm\sqrt{3}}{3}, i = 1, ..., n\}$. Also, the Hessian matrix of this function is given by,

$$\begin{pmatrix} 6(x_1 - 1) & 0 & \cdots & 0 \\ 0 & 6(x_2 - 1) & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & \cdots & \cdots & 6(x_n - 1) \end{pmatrix}$$

Hence, there are one local maximum, one local minimum, and $(2^n - 2)$ saddle points.

3. (No free lunch theorem) *You are given a binary random variable X and are told to estimate the value of another binary random variable Y. There can be many learning algorithms producing an estimate $\hat{Y}$ of Y based on the observation X, e.g., $\hat{Y} = 0$, $\hat{Y} = 1$, $\hat{Y} = X$, or $\hat{Y} = 1 - X$, i.e., the output of the algorithm is "always 0", "always 1", "same as the first outcome", and "different from the first outcome", respectively. There can also be randomized algorithms, e.g., set $\hat{Y} = 1$ randomly with probability p or set $\hat{Y} = 1$ with probability p and set $\hat{Y} = X$ with probability $1 - p$.*

(a) *Assume X and Y are i.i.d. Bernoulli(1/2). Show that no learning algorithm can perform any better than random guessing, i.e., producing $\hat{Y}$ as Bernoulli(1/2) regardless of X.*

Solution

In this setting, we can express any learning algorithm as

$$\Pr(\hat{Y} = 1|X = x) = \begin{cases} p & \text{if} \quad x = 0 \\ q & \text{if} \quad x = 1, \end{cases} \tag{1}$$

where $0 \leq p, q \leq 1$. Then, the probability of error of the learning algorithm is given by

$$\sum_{(x,y)\in\{0,1\}^2} \Pr(X = x, Y = y) \Pr(\hat{Y} \neq y|X = x) \tag{2}$$

We want to find p and q that minimize the error probability.

When X and Y are i.i.d. Bernoulli(1/2), the error probability becomes

$$\sum_{(x,y)\in\{0,1\}^2} \frac{1}{4} \Pr(\hat{Y} \neq y|X = x) = \frac{1}{2} \tag{3}$$

Since this is the same for all p and q, we can say no learning algorithm can perform better than any other learning algorithm including random guessing.

(b) *Can there be another joint distribution $p(x, y)$ on (X, Y) such that no learning algorithm can perform any better than random guessing?*

Solution

We prove that there is no learning algorithm that outperforms the random guessing if and only if X and Y are independent and Y is Bernoulli($\frac{1}{2}$).

Consider X and Y whose joint probability is given by

$$\Pr(X = x, Y = y) = \begin{cases} a & \text{if } x = 0,\ y = 0 \\ b & \text{if } x = 0,\ y = 1 \\ c & \text{if } x = 1,\ y = 0 \\ d & \text{if } x = 1,\ y = 1 \end{cases} \tag{4}$$

where $0 \le a, 0 \le b, 0 \le c,\ 0 \le d,\ a + b + c + d = 1$.

Also, let's consider a learning algorithm defined as (1). Then, the error probability of the learning algorithm is

$$a \Pr(\hat{Y} \neq 0 | X = 0) + b \Pr(\hat{Y} \neq 1 | X = 0) + c \Pr(\hat{Y} \neq 0 | X = 1) + d \Pr(\hat{Y} \neq 1 | X = 1) \tag{5}$$
$$= ap + b(1 - p) + cq + d(1 - q) \tag{6}$$
$$= p(a - b) + q(c - d) + b + d \tag{7}$$

The condition that X and Y are independent and Y is Bernoulli(1/2) is equivalent to $a = b$ and $c = d$, which implies $a + c = b + d = \frac{1}{2}$. In this case (7) becomes $\frac{1}{2}$ for all p, q and no learning algorithm can perform better than any other algorithm including random guessing.

Now assume the condition does not hold, then either $a \neq b$ or $c \neq d$ holds. If $a \neq b$, then we can choose $q = \frac{1}{2}$ and $p = 0$ or $p = 1$ such that (7) is strictly smaller than $\frac{1}{2}$ because (7) is linear in p and becomes $\frac{1}{2}$ if random guessing is done, i.e., $p = q = \frac{1}{2}$. Similar conclusion holds when $c \neq d$.

Therefore, the only distribution $p(x, y)$ for which no learning algorithm can perform any better than random guessing is when X and Y are independent and Y is Bernoulli(1/2).

The "no free lunch" theorem by Wolpert and Macready in 1997 basically states that any two optimization algorithms perform the same if averaged over all possible problems." In this simple toy problem, we can try to see what it means by averaging (7) over all possible values of a, b, c, d. Namely, if we average (7) over the uniform distribution of (a, b, c, d) such that $a, b, c, d \ge 0$ and $a + b + c + d = 1$, then the error probability is given by

$$\frac{\int_0^1 \int_0^{1-a} \int_0^{1-a-b} ap + b(1 - p) + cq + (1 - a - b - c)(1 - q)\ \mathrm{d}c\, \mathrm{d}b\, \mathrm{d}a}{\int_0^1 \int_0^{1-a} \int_0^{1-a-b} \mathrm{d}c\, \mathrm{d}b\, \mathrm{d}a} = \frac{1}{2} \tag{8}$$

We can see this is independent of p and q.