Distribusi dan Algoritma Naive Bayes

1. Untuk menerapkan algoritma Naive Bayes, Anda membuat sebuah data sintetis yang merepresentasikan studi kasus yang ingin diselesaikan, yakni klasifikasi jenis burung. Manakah dari opsi berikut sebagai tahapan yang tepat dalam membuat data sintetis tersebut?
    a. Pertama, buat 5000 data acak untuk menyimulasikan distribusi binomial. Gunakan nilai-nilai acak tersebut dalam perhitungan berikut, ketika nilai *x* akan mengikuti bentuk distribusi binomial.

$$x = F^{-1}(y) = \sigma\sqrt{2} \cdot \operatorname{erf}^{-1}(2y - 1) + \mu$$

    b. Pertama, buatlah nilai acak yang mengikuti distribusi uniform dengan interval [0,1]. Selanjutnya, gunakan fungsi inverse dari CDF distribusi binomial dengan bantuan library, seperti `**numpy.random.binomial**` untuk mengubah nilai uniform tersebut menjadi angka yang mengikuti pola distribusi binomial. Nilai inverse tersebut adalah distribusi acak yang akan berbentuk seperti distribusi binomial.
    c. Mulailah dengan membuat data acak yang mengikuti distribusi Gaussian. Lalu, lakukan inverse CDF distribusi tersebut untuk memperoleh nilai-nilai dalam distribusi uniform. Dari distribusi uniform tersebut, gunakan rumus di bawah ini untuk menghasilkan data acak yang mengikuti distribusi binomial.

$$x = F^{-1}(y) = \binom{n}{k} p^k (1-p)^{n-k}$$

    d. Mulailah dari menghasilkan sejumlah nilai acak sesuai dengan kebutuhan. Selanjutnya, gunakan rumus CDF berikut untuk mengubah nilai-nilai tersebut menjadi data acak yang mengikuti distribusi binomial.

$$F(x) = P(X \leq x) = P(X = 0) + P(X = 1) + \ldots + P(X = \lfloor x \rfloor) = \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1-p)^{n-k}$$

2. Dalam studi kasus yang dihadapi, Anda membuat data sintetis berdasarkan konsep variabel acak. Jika suatu variabel acak diasumsikan dengan `$n \leq x \leq m$`, manakah opsi di bawah ini yang merupakan rumus untuk **menghitung distribusi uniform**-nya?
    a. $P(X = x) = 1/(m - n)$
    b. $P(X = x) = Px$
    c. $P(X = k) = \int_{n}^{m} f(x)dx$
    d. $P(X = k) = m/n$

3. Dalam penerapan algoritma Naive Bayes, Anda perlu menghitung nilai *likelihood* yang merupakan bagian dari algoritmanya. Bagaimana cara menghitung nilai *likelihood* dalam algoritma Naive Bayes jika data atau atribut yang digunakan bertipe kontinu?
    a. Menghitung frekuensi setiap nilai untuk masing-masing kelas target.
    b. Menghitung frekuensi nilai yang paling sering dan jarang muncul pada setiap kelas.
    c. Mengasumsikan nilai-nilai kontinu mengikuti suatu pola distribusi, lalu menghitung *cumulative distribution function (CDF)* berdasarkan data yang tersedia.
    d. Mengasumsikan bahwa nilai-nilai kontinu mengikuti distribusi tertentu, lalu menghitungnya dengan *probability density function (PDF)* dari distribusi tersebut.
4. Dalam soal yang dikerjakan, Anda diminta melakukan perhitungan distribusi binomial menggunakan probability mass function (PMF). Mengapa pendekatan ini yang digunakan?
    a. Karena distribusi binomial merupakan distribusi diskret.
    b. Karena distribusi binomial hanya menyimpan nilai-nilai kategorikal.
    c. Karena distribusi ini hanya bisa digunakan untuk *probability mass function* (PMF).
    d. Karena nilai yang disimpan berada pada interval [0,1].
5. Dalam kasus penerapan algoritma Naive Bayes, apakah **alasan paling tepat** yang menyatakan bahwa kolom **sing_days** pada data sintetis harus berdistribusi binomial?
    a. Data yang tersimpan berupa jumlah kicauan burung bersifat diskret/kontinu.
    b. Data yang tersimpan berupa jumlah hari burung berkicau pada kolom ini bersifat kategorikal.
    c. Data yang tersimpan pada kolom ini memodelkan jumlah hari dalam sebulan, yakni seekor burung berhasil berkicau.
    d. Data jumlah hari burung berkicau pada kolom ini memodelkan nilai acak dalam interval [0,1] pada satuan bulan.