Machine Learning (Bayesian Classifier for continuous Attributes)

Prof K R Chowdhary

MBM University

October 18, 2024

Continuous Attribute Vectors

⇒ In continuous attributes, e.g, age, percentile marks, speed of car, temperature values, relative frequency is impractical.

⇒ Let a population of 900, we want to find, given one person as sample, to what age he/she belongs? But, there can be infinite number of ages. So, we divide ages in 10 intervals $(0, 10]$, ..., $(90, 100]$, like 10 different attributes.

⇒ Frequency count in age interval a_1 to a_{10} represented by '$\times$' signs, one '$\times$' is population of 30. Population at $(30, 40]$ (a_4) is $4 \times 30 = 120$.

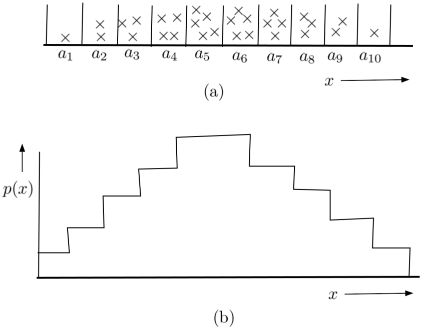

Figure 1: (a) Population at each age intervals, (b) Histogram plot of age interval x versus population density $p(x)$

⇒ In histogram, each step i corresponds to population count in interval a_i. A function $p(x)$, has value in ith slot as N_i/N. So, $\frac{\sum N_i}{N} = 1$.

⇒ We may shorten the interval in the histogram by increasing count of intervals, and ensure that number of persons in each slot are sufficient for reliable probability estimates.

⇒ In a general case, we keep reducing the length of interval until it becomes infinitesimally small, plot (Fig. 1) becomes a continuous function $p(x)$ (Fig. 2). High and low count refers to density of people, so $p(x)$ is *probability density function* (pdf).

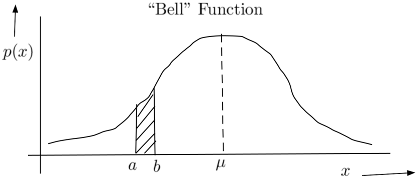

Figure 2: Probability Density Function (Bell curve)

⇒ In Fig. 2, slot "a to b" is probability of $x \in [a, b]$. It is relative size of the area under this section of pdf curve.

⇒ $p(x)$ is probability at x. If pdf has been created exclusively from examples in class c_i then this probability is $p_{c_i}(x)$, in discrete attributes it was $P(\mathrm{x}|c_i)$.

Bayes Formula for Continuous Attributes: In pdf it is possible to use Bayes formula. Now, conditional probability $P(x|c_i)$ becomes $p_{c_i}(x)$, and $P(x)$ becomes $p(x)$.

For single attribute x, Bayes formula is:

$$P(c_i|x) = \frac{p_{c_i}(x).P(c_i)}{p(x)}, \quad (1)$$

⇒ $P(c_i)$ = Estimated relative freq. of class c_i in training set, $p(x)$ = pdf due to all training examples, $p_{c_i}(x)$ = pdf due to training examples in class c_i.

⇒ We assume attributes as mutually exclusive. For a vector $\mathbf{x} = \{x_1, ..., x_n\}$, pdf is:

$$p_{c_j}(\mathbf{x}) = \prod_{i=1}^{n} p_{c_j}(x_i) \quad (2)$$

⇒ After discritizing continuous attribute we get approximate pdf. We can also use standard probability model, known as Gaussian function. Shape of the function is "bell" function (Fig. 2), maximum is at $x = \mu$ (*mean*), towards both directions height decreases. Gaussian curve can be represented by:

$$p(x) = k.e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad (3)$$

⇒ The e is base of natural log, σ is *variance*. Greater is the difference between x and μ, smaller will be $p(x)$. How steep is slop, depends on σ^2. Greater variance means smaller sensitivity to the difference between μ and x, and it will result to a flatter bell curve. Smaller value of σ, it will result to a narrower bell curve.

⇒ Coefficient k makes the area under the curve as 1, which is a requirement for the theory of probability. This happens when value of k is,

$$k = \frac{1}{\sqrt{2\pi\sigma^2}}. \qquad (4)$$

Area under the Bell curve is unity

To show that area under bell curve for pdf $p(x) = ke^{-\frac{(x-\mu)^2}{2\sigma^2}}$ is 1, we substitute value of k:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Next, we need to calculate the integral of $p(x)$ over the entire range of x, which is from $-\infty$ to $+\infty$:

$$\int_{-\infty}^{+\infty} p(x)dx = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

To solve this integral, we can make a substitution. Let: $z = \frac{x-\mu}{\sigma}$. Taking derivative both sides, we get, $dx = \sigma dz$. The limits of integration remain the same as x approaches $-\infty$ and $+\infty$. When $x = -\infty$, then $z = -\infty$ and when $x = +\infty$, $z = +\infty$.

Area under the Bell curve is unity ...

We can rewrite the integral:

$$\int_{-\infty}^{+\infty} p(x)dx = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z\sigma)^2}{2\sigma^2}} \sigma dz$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \sigma \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz$$

The integral $\int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz$ is a well-known result, and evaluates to:

$$\int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz = \sqrt{2\pi}$$

Substituting this back into our expression, we get:

$$\int_{-\infty}^{+\infty} p(x)dx = \frac{1}{\sqrt{2\pi\sigma^2}} \sigma \cdot \sqrt{2\pi} = 1$$

which shows that area under the bell curve is 1, and $p(x)$ is a valid probability density function.

Parameter Values

⇒ Since formula (3) is standard Bell curve, we can use it to calculate probability $p_{c_i}(x)$. For this we require μ and σ. Let there are m classes of c_i in training set, x_i is value of given attribute in i-th example, then mean (μ) and variance (σ) are:

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i, \qquad (5)$$

and

$$\sigma^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \mu)^2. \qquad (6)$$

⇒ Centre of Gaussian curve (i.e., μ) is obtained by arithmetic average of values observed in training examples, and variance (σ) is obtained by squaring difference of x_i and μ.

⇒ To calculate the variance, we divide expression by $m - 1$ and not by m, this is to compensate that, μ itself is an estimate. The variance should therefore be some what higher then it would be if we divided by m.