# Gaussian naive Bayes classifier

**DEFINITION 1:**

Sample size: $N$

Sample mean:

$$\mu = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i$$

Unbiased sample variance:

$$\sigma^2 = \frac{1}{N-1} \cdot \sum_{i=1}^{N} (x_i - \mu)^2$$

**TRAINING:**

Training set example:

| **Sex (class)** | **Height (feet)** | **Weight (lbs)** | **Foot size (inches)** |
|---|---|---|---|
| Male | 6 | 180 | 12 |
| Male | 5.92 | 190 | 11 |
| Male | 5.58 | 170 | 12 |
| Male | 5.92 | 165 | 10 |
| Female | 5 | 100 | 6 |
| Female | 5.5 | 150 | 8 |
| Female | 5.42 | 130 | 7 |
| Female | 5.75 | 150 | 9 |

Computing means and variances of example training set:

$$\mu_{Height\ Male} = \frac{1}{4} \cdot (6 + 5.92 + 5.58 + 5.92) = \mathbf{5.855}$$

$$\sigma^2_{Height\ Male} = \frac{1}{3} \cdot ((6 - 5.855)^2 + (5.92 - 5.855)^2 + (5.58 - 5.855)^2 + (5.92 - 5.855)^2) \approx \mathbf{0.035033}$$

$$\mu_{Height\ Female} = \frac{1}{4} \cdot (5 + 5.5 + 5.42 + 5.75) = \mathbf{5.4175}$$

$$\sigma^2_{Height\ Female} = \frac{1}{3} \cdot ((5 - 5.4175)^2 + (5.5 - 5.4175)^2 + (5.42 - 5.4175)^2 + (5.75 - 5.4175)^2) = \mathbf{0.097225}$$

$$\mu_{Weight\ Male} = \frac{1}{4} \cdot (180 + 190 + 170 + 165) = \mathbf{176.25}$$

$$\sigma^2_{Weight\ Male} = \frac{1}{3} \cdot ((180 - 176.25)^2 + (190 - 176.25)^2 + (170 - 176.25)^2 + (165 - 176.25)^2) \approx \mathbf{122.92}$$

$$\mu_{Weight\ Female} = \frac{1}{4} \cdot (100 + 150 + 130 + 150) = \mathbf{132.5}$$

$$\sigma^2_{Weight\ Female} = \frac{1}{3} \cdot ((100 - 132.5)^2 + (150 - 132.5)^2 + (130 - 132.5)^2 + (150 - 132.5)^2) \approx \mathbf{558.33}$$

$$\mu_{Foot\ Size\ Male} = \frac{1}{4} \cdot (12 + 11 + 12 + 10) = \mathbf{11.25}$$

$$\sigma^2_{Foot\ Size\ Male} = \frac{1}{3} \cdot ((12 - 11.25)^2 + (11 - 11.25)^2 + (12 - 11.25)^2 + (10 - 11.25)^2) \approx \mathbf{0.9167}$$

$$\mu_{Foot\ Size\ Female} = \frac{1}{4} \cdot (6 + 8 + 7 + 9) = \mathbf{7.5}$$

$$\sigma^2_{Foot\ Size\ Female} = \frac{1}{3} \cdot ((6 - 7.5)^2 + (8 - 7.5)^2 + (7 - 7.5)^2 + (9 - 7.5)^2) \approx \mathbf{1.6667}$$

| **Sex (class)** | **Mean (Height)** | **Variance (Height)** | **Mean (Weight)** | **Variance (Weight)** | **Mean (Foot size)** | **Variance (Foot size)** |
|---|---|---|---|---|---|---|
| Male | 5.855 | 0.035033 | 176.25 | 122.92 | 11.25 | 0.9167 |
| Female | 5.4175 | 0.097225 | 132.5 | 558.33 | 7.5 | 1.6667 |

**DEFINITION 2:**

Probability of class: $P(Male)$, $P(Female)$

In our case classes are equiprobable: $P(Male) = P(Female) = 0.5$

Probability density function:

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2 \cdot \sigma^2}}$$

Posterior probability male:

$$posterior(male) = \frac{P(male) \cdot p(height|male) \cdot p(weight|male) \cdot p(foot\ size|male)}{evidence}$$

Posterior probability female:

$$posterior(female) = \frac{P(female) \cdot p(height|female) \cdot p(weight|female) \cdot p(foot\ size|female)}{evidence}$$

Evidence:

$$evidence = P(male) \cdot p(height|male) \cdot p(weight|male) \cdot p(foot\ size|male) + P(female) \cdot p(height|female) \cdot p(weight|female) \cdot p(foot\ size|female)$$

**TESTING**

Testing sample:

| Sex (class) | Height (feet) | Weight (lbs) | Foot size (inches) |
|---|---|---|---|
| ? | 6 | 130 | 8 |

Computing probability densities:

$$p(height|male) = \frac{1}{\sqrt{2 \cdot \pi \cdot 0.035033}} \cdot e^{-\frac{(6-5.855)^2}{2 \cdot 0.035033}} \approx \mathbf{1.5788}$$

$$p(weight|male) = \frac{1}{\sqrt{2 \cdot \pi \cdot 122.92}} \cdot e^{-\frac{(130-176.25)^2}{2 \cdot 122.92}} \approx \mathbf{0.0000059881}$$

$$p(foot\ size|male) = \frac{1}{\sqrt{2 \cdot \pi \cdot 0.9167}} \cdot e^{-\frac{(8-11.25)^2}{2 \cdot 0.9167}} \approx \mathbf{0.0013115}$$

$$p(height|female) = \frac{1}{\sqrt{2 \cdot \pi \cdot 0.097225}} \cdot e^{-\frac{(6-5.4175)^2}{2 \cdot 0.097225}} \approx \mathbf{0.22346}$$

$$p(weight|female) = \frac{1}{\sqrt{2 \cdot \pi \cdot 558.33}} \cdot e^{-\frac{(130-132.5)^2}{2 \cdot 558.33}} \approx \mathbf{0.016789}$$

$$p(foot\ size|female) = \frac{1}{\sqrt{2 \cdot \pi \cdot 1.6667}} \cdot e^{-\frac{(8-7.5)^2}{2 \cdot 1.6667}} \approx \mathbf{0.28669}$$

$$P(male) \cdot p(height|male) \cdot p(weight|male) \cdot p(foot\ size|male) \approx 0.5 \cdot 1.5788 \cdot 0.0000059881 \cdot 0.0013115 \approx 6.19947 \cdot 10^{-9}$$

$$P(female) \cdot p(height|female) \cdot p(weight|female) \cdot p(foot\ size|female) \approx 0.5 \cdot 0.22346 \cdot 0.016789 \cdot 0.28669 \approx 5.37783 \cdot 10^{-4}$$

Since $5.37783 \cdot 10^{-4} > 6.19947 \cdot 10^{-9}$, we predict the sample is female.

**REFERENCES:**

http://en.wikipedia.org/wiki/Naive_Bayes_classifier

http://guidetodatamining.com/guide/ch6/DataMining-ch6.pdf