

CMP-6002B - Machine Learning

Dr Gavin Cawley

Seminar 1 : Basic Principles

Question

A medical screening test, which measures the level of a particular enzyme, is available for a serious disease known as “Bagnall’s syndrome”¹. Table 1 shows the raw enzyme levels for a sample of patients that are either confirmed cases of Bagnall’s syndrome or are known to be free of this particular disease. Construct a parametric classifier to predict whether a patient is more likely than not to be suffering from Bagnall’s syndrome.

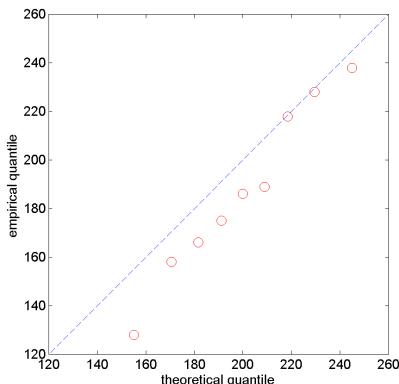
Table 2. Calibration data

Disease status	Measured enzyme level (ppm)				
positive	128	175	186	228	158
	166	218	241	238	189
negative	96	124	106	81	90
	85	76	106	154	102

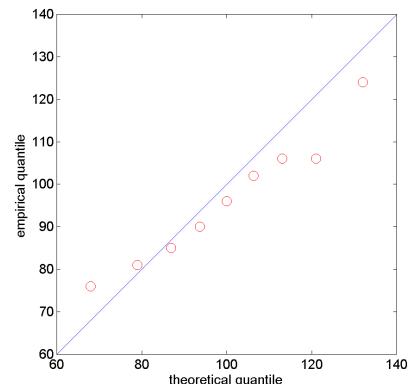
¹Symptoms: Obsession with Arsenal FC.

Distributional Assumptions

- ▶ Are the data normally distributed?
 - ▶ View quantile-quantile (q-q) plots
 - ▶ Straight diagonal line at 45° means *exactly* normal



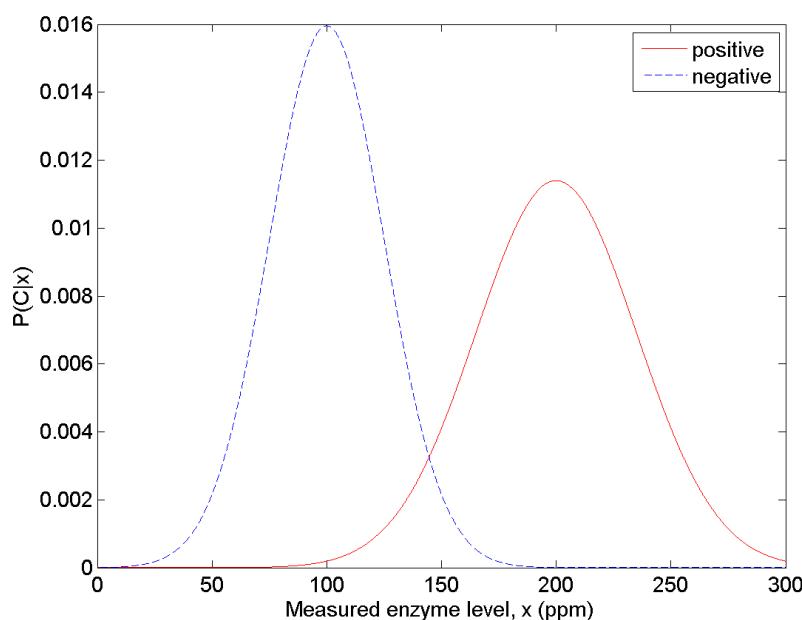
positive patterns



negative patterns

- ▶ Good enough (not much data so plots will be a bit noisy)!

True Class Conditional Densities



- ▶ I generated the data from Gaussians!

Fitting the model

- ▶ Adopt a Gaussian likelihood for each class,

$$p(x|\mathcal{C}_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{-\frac{(x - \mu_k)^2}{2\sigma_k^2}\right\}.$$

- ▶ Distributions characterised by a mean μ_k and a variance, σ_k^2 .

$$\begin{aligned}\mu_k &= \frac{1}{N_k} \sum_{i \in \mathcal{C}_k} x_i, & \mu_p &= 192.7 \\ \sigma_k^2 &= \frac{1}{N_k - 1} \sum_{i \in \mathcal{C}_k} (x_i - \mu_k)^2. & \sigma_p &= 37.6329 \\ && \mu_n &= 102.0 \\ && \sigma_n &= 23.1084\end{aligned}$$

- ▶ The prior probability, $p(\mathcal{C}_k) = N_k/N$.

$$p(\mathcal{C}_p) = \frac{10}{20} = 0.5 \quad p(\mathcal{C}_n) = \frac{10}{20} = 0.5$$

Making Predictions

- ▶ Use Bayes' rule:

$$p(\mathcal{C}_k|x) = \frac{p(x|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_{i=1}^c p(x|\mathcal{C}_i)p(\mathcal{C}_i)}$$

Table 2. Patients

Patient	Measured enzyme level (ppm)
Tom	171
Bert	123
Ernie	156

- ▶ Compute the probability that each patient is suffering from Bagnall's Syndrome

Tom - Measured Enzyme Level = 171 ppm

- ▶ Compute likelihoods

$$\begin{aligned}
 p(x = 171 | \mathcal{C}_p) &= \frac{1}{\sqrt{2\pi} \times 37.6329} \exp \left\{ -\frac{(171 - 192.7)^2}{2 \times 37.6329^2} \right\} \\
 &= 0.0090 \\
 p(x = 171 | \mathcal{C}_n) &= \frac{1}{\sqrt{2\pi} \times 23.1084} \exp \left\{ -\frac{(171 - 102.0)^2}{2 \times 23.1084^2} \right\} \\
 &= 2.1887E^{-4}
 \end{aligned}$$

- ▶ Apply Bayes' rule

$$\begin{aligned}
 p(\mathcal{C}_p | x = 171) &= \frac{p(x = 171 | \mathcal{C}_p)p(\mathcal{C}_p)}{p(x = 171 | \mathcal{C}_p)p(\mathcal{C}_p) + p(x = 171 | \mathcal{C}_n)p(\mathcal{C}_n)} \\
 &= \frac{0.0090 \times 0.5}{0.0090 \times 0.5 + 2.1887E^{-4} \times 0.5} = 0.9762 \\
 p(\mathcal{C}_n | x = 171) &= 1 - p(\mathcal{C}_p | x = 171) = 0.0238
 \end{aligned}$$

Bert - Measured Enzyme Level = 123 ppm

- ▶ Compute likelihoods

$$\begin{aligned}
 p(x = 123 | \mathcal{C}_p) &= \frac{1}{\sqrt{2\pi} \times 37.6329} \exp \left\{ -\frac{(123 - 192.7)^2}{2 \times 37.6329^2} \right\} \\
 &= 0.0019 \\
 p(x = 123 | \mathcal{C}_n) &= \frac{1}{\sqrt{2\pi} \times 23.1084} \exp \left\{ -\frac{(123 - 102.0)^2}{2 \times 23.1084^2} \right\} \\
 &= 0.0117
 \end{aligned}$$

- ▶ Apply Bayes' rule

$$\begin{aligned}
 p(\mathcal{C}_p | x = 123) &= \frac{p(x = 123 | \mathcal{C}_p)p(\mathcal{C}_p)}{p(x = 123 | \mathcal{C}_p)p(\mathcal{C}_p) + p(x = 123 | \mathcal{C}_n)p(\mathcal{C}_n)} \\
 &= \frac{0.0019 \times 0.5}{0.0019 \times 0.5 + 0.0117 \times 0.5} = 0.1398 \\
 p(\mathcal{C}_n | x = 123) &= 1 - p(\mathcal{C}_p | x = 123) = 0.8602
 \end{aligned}$$

Ernie - Measured Enzyme Level = 156 ppm

- ▶ Compute likelihoods

$$\begin{aligned} p(x = 156|\mathcal{C}_p) &= \frac{1}{\sqrt{2\pi} \times 37.6329} \exp \left\{ -\frac{(156 - 192.7)^2}{2 \times 37.6329^2} \right\} \\ &= 0.0066 \\ p(x = 156|\mathcal{C}_n) &= \frac{1}{\sqrt{2\pi} \times 23.1084} \exp \left\{ -\frac{(156 - 102.0)^2}{2 \times 23.1084^2} \right\} \\ &= 0.0012 \end{aligned}$$

- ▶ Apply Bayes' rule

$$\begin{aligned} p(\mathcal{C}_p|x = 156) &= \frac{p(x = 156|\mathcal{C}_p)p(\mathcal{C}_p)}{p(x = 156|\mathcal{C}_p)p(\mathcal{C}_p) + p(x = 156|\mathcal{C}_n)p(\mathcal{C}_n)} \\ &= \frac{0.0066 \times 0.5}{0.0066 \times 0.5 + 0.0012 \times 0.5} = 0.8451 \\ p(\mathcal{C}_n|x = 156) &= 1 - p(\mathcal{C}_p|x = 156) = 0.1549 \end{aligned}$$

Dealing with Unequal Operational Priors

- ▶ The training data has been artificially “balanced”
 - ▶ Training set priors: $p(\mathcal{C}_p) = 0.5$ and $p(\mathcal{C}_n) = 0.5$
- ▶ Now we are told that the proportion of people suffering from Bagnall's syndrome in the general public is one in 1000. Recompute the probabilities that Tom, Bert and Ernie really are suffering from Bagnall's syndrome.
 - ▶ Operational priors: $p_o(\mathcal{C}_p) = 0.001$ and $p_o(\mathcal{C}_n) = 0.999$
- ▶ Use Bayes' rule, $p(\mathcal{C}_k|x) \propto p(x|\mathcal{C}_k)p(\mathcal{C}_k)$, so

$$p(x|\mathcal{C}_k)p(\mathcal{C}_k) \times \frac{p_o(\mathcal{C}_k)}{p(\mathcal{C}_k)} = p(x|\mathcal{C}_k)p_o(\mathcal{C}_k) = q_o(\mathcal{C}_k|x) \propto p_o(\mathcal{C}_k|x)$$

- ▶ Multiply by ratio of prior probabilities, then normalise so that probabilities sum to one.

Tom, measured enzyme level = 171 ppm

- Step 1 - scale:

$$\begin{aligned} q_o(\mathcal{C}_p|x=171) &= p(\mathcal{C}_p|x=171) \frac{p_o(\mathcal{C}_p)}{p(\mathcal{C}_p)} \\ &= 0.9762 \times \frac{0.001}{0.5} = 0.0020 \\ q_o(\mathcal{C}_n|x=171) &= p(\mathcal{C}_n|x=171) \frac{p_o(\mathcal{C}_n)}{p(\mathcal{C}_n)} \\ &= 0.0238 \times \frac{0.999}{0.5} = 0.0476 \end{aligned}$$

- Step 2 - Normalise:

$$\begin{aligned} p_o(\mathcal{C}_p|x=171) &= \frac{q_o(\mathcal{C}_p|x=171)}{q_o(\mathcal{C}_p|x=171) + q_o(\mathcal{C}_n|x=171)} \\ &= \frac{0.0020}{0.0020 + 0.0476} = 0.0403 \\ p_o(\mathcal{C}_n|x=171) &= 1 - p_o(\mathcal{C}_p) = 1 - 0.0403 = 0.9597 \end{aligned}$$

Bert, measured enzyme level = 123 ppm

- Step 1 - scale:

$$\begin{aligned} q_o(\mathcal{C}_p|x=123) &= p(\mathcal{C}_p|x=123) \frac{p_o(\mathcal{C}_p)}{p(\mathcal{C}_p)} \\ &= 0.1398 \times \frac{0.001}{0.5} = 2.7960E^{-4} \\ q_o(\mathcal{C}_n|x=123) &= p(\mathcal{C}_n|x=123) \frac{p_o(\mathcal{C}_n)}{p(\mathcal{C}_n)} \\ &= 0.8602 \times \frac{0.999}{0.5} = 1.7187 \end{aligned}$$

- Step 2 - Normalise:

$$\begin{aligned} p_o(\mathcal{C}_p|x=123) &= \frac{q_o(\mathcal{C}_p|x=123)}{q_o(\mathcal{C}_p|x=123) + q_o(\mathcal{C}_n|x=123)} \\ &= \frac{2.7960E^{-4}}{2.796E^{-4} + 1.7187} = 1.6265E^{-4} \\ p_o(\mathcal{C}_n|x=123) &= 1 - p_o(\mathcal{C}_p) = 1 - 1.6265E^{-4} = 0.9998 \end{aligned}$$

Ernie, measured enzyme level = 156 ppm

- ▶ Step 1 - scale:

$$\begin{aligned} q_o(\mathcal{C}_p|x=156) &= p(\mathcal{C}_p|x=156) \frac{p_o(\mathcal{C}_p)}{p(\mathcal{C}_p)} \\ &= 0.8451 \times \frac{0.001}{0.5} = 0.0017 \\ q_o(\mathcal{C}_n|x=156) &= p(\mathcal{C}_n|x=156) \frac{p_o(\mathcal{C}_n)}{p(\mathcal{C}_n)} \\ &= 0.1459 \times \frac{0.999}{0.5} = 0.2915 \end{aligned}$$

- ▶ Step 2 - Normalise:

$$\begin{aligned} p_o(\mathcal{C}_p|x=156) &= \frac{q_o(\mathcal{C}_p|x=156)}{q_o(\mathcal{C}_p|x=156) + q_o(\mathcal{C}_n|x=156)} \\ &= \frac{0.0017}{0.0017 + 0.2915} = 0.0058 \\ p_o(\mathcal{C}_n|x=156) &= 1 - p_o(\mathcal{C}_p) = 1 - 0.0058 = 0.9942 \end{aligned}$$

Making Minimum Risk Decisions

- ▶ We are now told that Bagnall's syndrome is easily cured at a cost of 1 per patient (the treatment has no side-effects), but the cost of an Arsenal season ticket is 1200. Should Tom, Bert and Ernie opt for treatment?
 - ▶ False positive cost = £1
 - ▶ False negative cost = £1199
- ▶ Expected loss associated with each decision:
 - ▶ Loss (treatment) = False positive cost $\times p_o(\mathcal{C}_n|x)$
 - ▶ Loss (no treatment) = False negative cost $\times p_o(\mathcal{C}_p|x)$
- ▶ Make decision with lowest expected loss.

Making Minimum Risk Decisions - Tom

- ▶ Tom: measured enzyme level = 171 (ppm)

$$\begin{aligned} p_o(\mathcal{C}_p|x=171) &= 0.043 \\ p_o(\mathcal{C}_n|x=171) &= 0.9597 \end{aligned}$$

- ▶ Expected loss for treatment:

$$L_{\text{treatment}} = p_o(\mathcal{C}_n|x=171) * C_{\text{fp}} = 0.9597 \times £1 = £0.96$$

- ▶ Expected loss for no treatment:

$$L_{\text{no treatment}} = p_o(\mathcal{C}_p|x=171) * C_{\text{fn}} = 0.043 \times £1199 = £51.60$$

- ▶ Minimum risk decision: treatment

Making Minimum Risk Decisions - Bert

- ▶ Bert: measured enzyme level = 123 (ppm)

$$\begin{aligned} p_o(\mathcal{C}_p|x=123) &= 1.6265E^{-4} \\ p_o(\mathcal{C}_n|x=123) &= 0.9998 \end{aligned}$$

- ▶ Expected loss for treatment:

$$L_{\text{treatment}} = p_o(\mathcal{C}_n|x=123) * C_{\text{fp}} = 0.9998 \times £1 = £0.9998$$

- ▶ Expected loss for no treatment:

$$L_{\text{no treatment}} = p_o(\mathcal{C}_p|x=123) * C_{\text{fn}} = 1.6265E^{-4} \times £1199 = £0.14$$

- ▶ Minimum risk decision: no treatment

Making Minimum Risk Decisions - Ernie

- ▶ Tom: measured enzyme level = 156 (ppm)

$$\begin{aligned} p_o(\mathcal{C}_p|x=156) &= 0.0058 \\ p_o(\mathcal{C}_n|x=156) &= 0.9942 \end{aligned}$$

- ▶ Expected loss for treatment:

$$L_{\text{treatment}} = p_o(\mathcal{C}_n|x=156) * C_{\text{fp}} = 0.9942 \times £1 = £0.99$$

- ▶ Expected loss for no treatment:

$$L_{\text{no treatment}} = p_o(\mathcal{C}_p|x=156) * C_{\text{fn}} = 0.0058 \times £1199 = £6.95$$

- ▶ Minimum risk decision: treatment

Conclusions

- ▶ Example of parametric statistical pattern recognition
- ▶ Dealing with unequal training set and operational priors
 - ▶ No need to retrain a probabilistic classifier
 - ▶ Scale and re-normalise
- ▶ Minimum risk decision making
 - ▶ The optimal decision depends on misclassification costs
 - ▶ A probabilistic classifier is required