Statistical Pattern Recognition

Andreas C. Kapourani

(Credit: Hiroshi Shimodaira)

1 Statistical Pattern Recognition

In many real life problems, we have to make decisions based on uncertainty, e.g. due to inaccurate or incomplete information about a problem. The mathematics of probability provides the way to deal with uncertainty, and tells us how to update our knowledge and beliefs if new information becomes available. In this lab session we introduce the use of probability and statistics for pattern recognition and learning.

Consider a pattern classification problem in which there are K classes. Let C denote the class, taking values $1, ..., K$, and let $P(C_k)$ be the prior probability of class k. The observed input data, which is a D-dimensional feature vector, is denoted by $\mathbf{X}$. Once the training set is used to train the classifier, a new, unlabelled data point $\mathbf{x}$ is observed. Let $P(\mathbf{x}|C_k)$ be the likelihood of class k for the data $\mathbf{x}$.

To perform classification we could use Bayes' theorem to compute the posterior probabilities $P(C_k|\mathbf{x})$, for every class $k = 1, ..., K$; we can then classify $\mathbf{x}$ by assigning it to the class with the highest posterior probability. That is, we need to compute:

$$P(C_k|\mathbf{x}) = \frac{P(\mathbf{x}|C_k)P(C_k)}{P(\mathbf{x})} \tag{1}$$

$$= \frac{P(\mathbf{x}|C_k)P(C_k)}{\sum_k P(\mathbf{x}|C_k)P(C_k)} \tag{2}$$

for every class k and then assign $\mathbf{x}$ to the class with the highest posterior probability (i.e. find $\arg\max_k P(C_k|\mathbf{x})$). This procedure is sometimes called MAP (*maximum a posteriori*) decision rule. Thus, for each class k we need to provide an estimate of the likelihood $P(\mathbf{x}|C_k)$ and the prior probability $P(C_k)$.

1.1 Fish Example data

To illustrate the use of posterior probabilities to perform classification, we will use a dataset which contains measurements of fish lengths. The dataset comprises 200 observations (100 male and 100 female fish), each representing the length of the fish. The objective is to classify the fish as *male* or *female* based on their length measurement. You can download the `Fish` dataset from the course website:

`http://www.inf.ed.ac.uk/teaching/courses/inf2b/learnLabSchedule.html`

You will find a file named `fish.txt`, download it and save it in your current folder. Note that this file is already pre-processed, and each line corresponds to three columns, the first column denotes the fish length x, the second and the third columns denote the number of male $n_M(x)$ and female $n_F(x)$ observations for that length, respectively.

Exercise

Read the file and load the data in MATLAB. Store the fish data in a matrix A.

1.2 Compute prior and likelihood

Let class $C = M$ represent male, and $C = F$ represent female fish. The *prior* probability expresses our beliefs about the sex of the fish before any evidence is taken into account. We can assume that male and female fish have different prior probabilities (e.g. $P(C_M) = 0.6, P(C_F) = 0.4$) or we can compute an estimate of those from the actual data by finding the proportion of male and female fish out of the total observations:

```
% Total number of male fish, i.e. 100
N_M = sum(A(:,2));

% Total number of female fish, i.e. 100
N_F = sum(A(:,3));

% total number of observations, i.e. 200
N_total = N_M + N_F;

% prior probability of male fish
prior_M = N_M / N_total
prior_M =
    0.5000
% prior probability for female is 1-P(M), since P(M) + P(F) = 1.
prior_F = 1 - prior_M
prior_F =
    0.5000
```

We can now estimate the *likelihoods* $P(x|C_M)$ and $P(x|C_F)$ as the counts in each class for length x divided by the total number of examples in that class:

$$P(x|C_M) \sim \frac{n_M(x)}{N_M} \tag{3}$$

$$P(x|C_F) \sim \frac{n_F(x)}{N_F} \tag{4}$$

Thus we can estimate the likelihoods of the length of each fish given each class using relative frequencies (i.e. using the training set of 100 examples from each class). Note that we obtain *estimates* of $P(x|C_M)$ and $P(x|C_F)$, since N_M and N_F are finite.

We can compute the likelihood for each fish length x, simply by computing the relative frequencies as follows:

```
% Likelihood vector for each length x for male fish
lik_M = A(:,2)/N_M;

% Likelihood vector for each length x for female fish
lik_F = A(:,3)/N_F;
```

Let's observe the length distribution for each class. We can do this easily by plotting histograms. Figure 1 shows the length distribution for male and female fish. For each class, also the Cumulative Distribution Function (CDF) is shown. The CDF is the probability that a real-valued random variable X will take a value less than or equal to x, that is, $CDF(x) = P(X \leq x)$, where P denotes the probability.

The code for creating plots (a) and (c) in Figure 1 is the following:

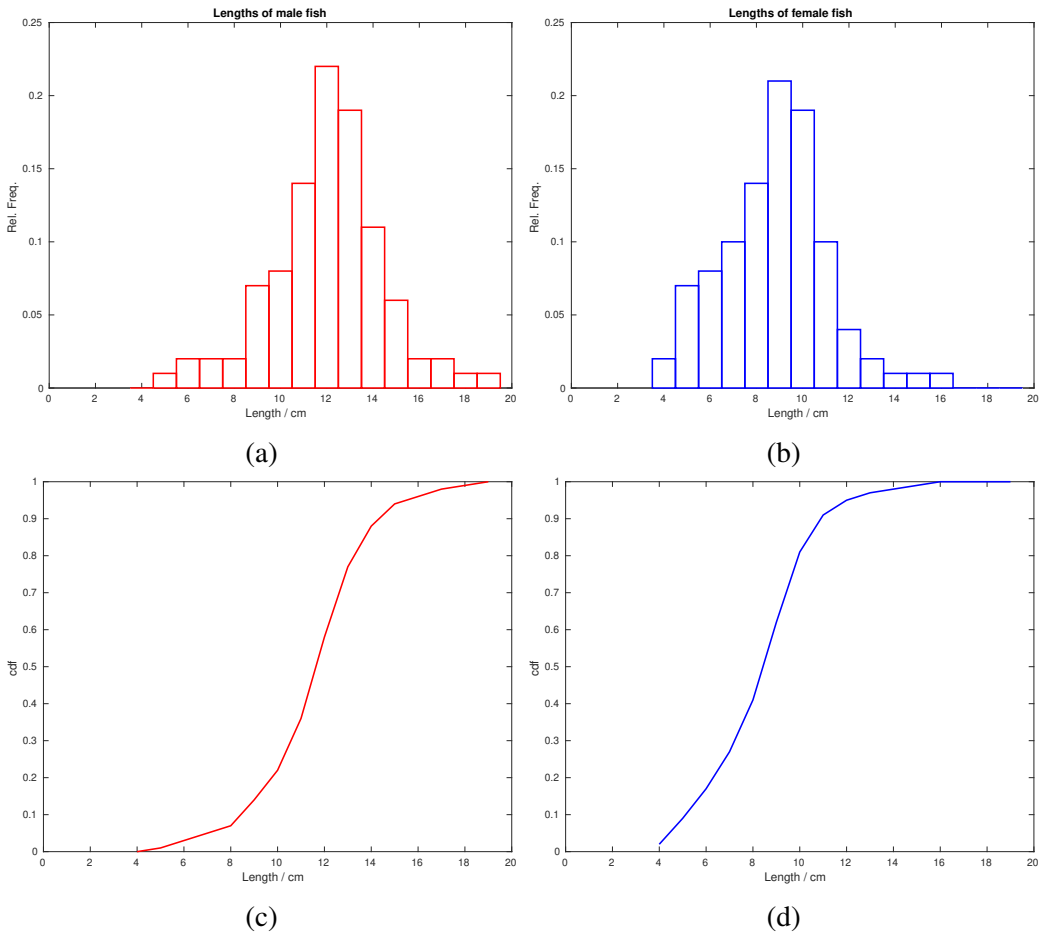

Figure 1: *(a) Relative frequency of lengths of male fish. (b) Relative frequency of lengths of female fish. (c) CDF for male fish. (d) CDF for female fish.*

```
% Create a histogram bar and return a vector of handles to this object
hh = bar(A(:,1), A(:,2)/N_M, 1);

% Modify the initial plot
set(hh, 'FaceColor', 'white');
set(hh, 'EdgeColor', 'red');
set(hh, 'linewidth', 1.5);
% Define x and y labels
ylabel('Rel. Freq.'); xlabel('Length / cm');
% Create title
title('Lengths of male fish');
% Define only x-axis limits
xlim([0 20]);

% Create CDF plot. Check what the 'cumsum' function does in Matlab
hh = plot(A(:,1), cumsum(A(:,2))/N_M, '-r');
% Modify the initial plot
set(hh, 'linewidth', 1.5);
% Define x and y labels
ylabel('cdf'); xlabel('Length / cm');
% Define only x-axis limits
```

```
xlim([0 20]);
```

We can also plot the likelihood $P(x|C_k)$ for each class as shown in Figure 2; note that the shapes of each class are similar to Figure 1, since we computed the likelihood from the relative frequencies. We observe that fish with length around $13cm$ are most likely to be male fish, since the likelihood is $P(x = 12|C_M) \approx 0.22$, whereas for female fish is only $P(x = 12|C_F) \approx 0.04$.

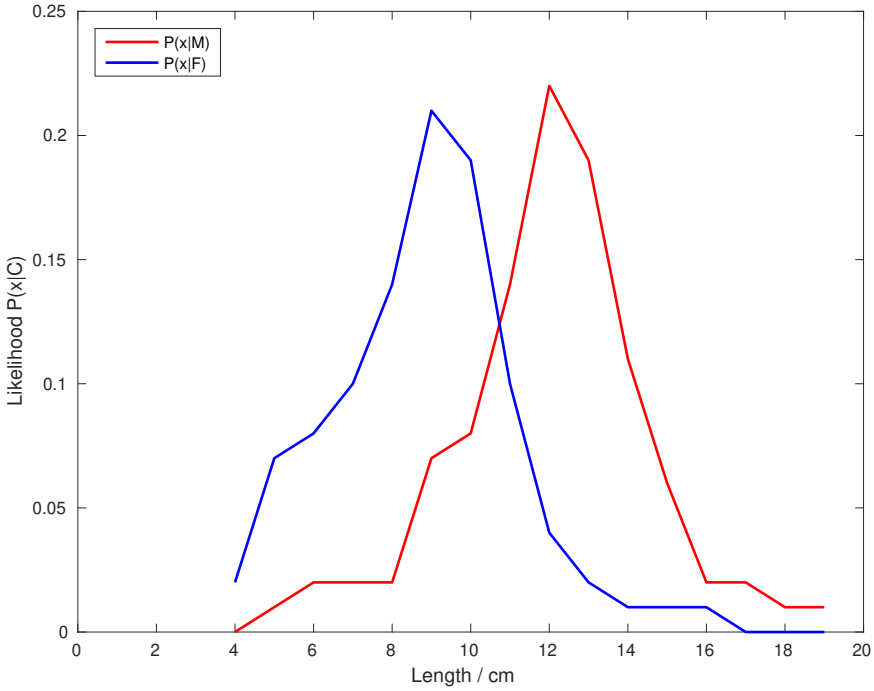

Figure 2: *Likelihood function for male and female fish lengths.*

Exercises

- Plot histogram of female fish lengths as shown in Figure 1 (b).
- Figures 1 (a) and (b), show relative frequencies. Modify your code so it can show the actual frequencies.
- Show both the male and female histograms in the same bar plot.
- Plot the likelihood functions for male and female fish lengths as shown in Figure 2.

1.3 Compute posterior probabilities

Having computed the prior and likelihood, we can now compute the posterior probabilities using Eq. 1. First we need to compute the *evidence* $P(\mathbf{x})$, which can be thought as a normalization constant ensuring that we have actual (posterior) probabilities (i.e. $0 \leq P(C_k|\mathbf{x}) \leq 1$ and $\sum_k P(C_k|\mathbf{x}) = 1$).

```
% Compute evidence vector for each fish length
Px = prior_M * lik_M + prior_F * lik_F;

% Compute vector of posterior probabilities for male fish lengths
post_M = lik_M * prior_M ./ Px;
```

```
% Compute vector of posterior probabilities for female fish lengths
post_F = lik_F * prior_F ./ Px;
```

We can now plot the posterior probabilities using the following code:

```
% Posterior probabilities for male fish
hh = plot(A(:,1), post_M, '-r');
set(hh, 'linewidth', 1.5);
ylabel('Posterior P(C|x)');
xlabel('Length / cm');
xlim([0 20]);

hold on
% Posterior probabilities for female fish
hh = plot(A(:,1), post_F, '-b');
set(hh, 'linewidth', 1.5);
legend('P(M|x)', 'P(F|x)', 'Location', 'northwest');

hold on
% Show decision boundary
dec_bound = 10.7;
plot([dec_bound dec_bound], get(gca, 'ylim'), '--k');
```

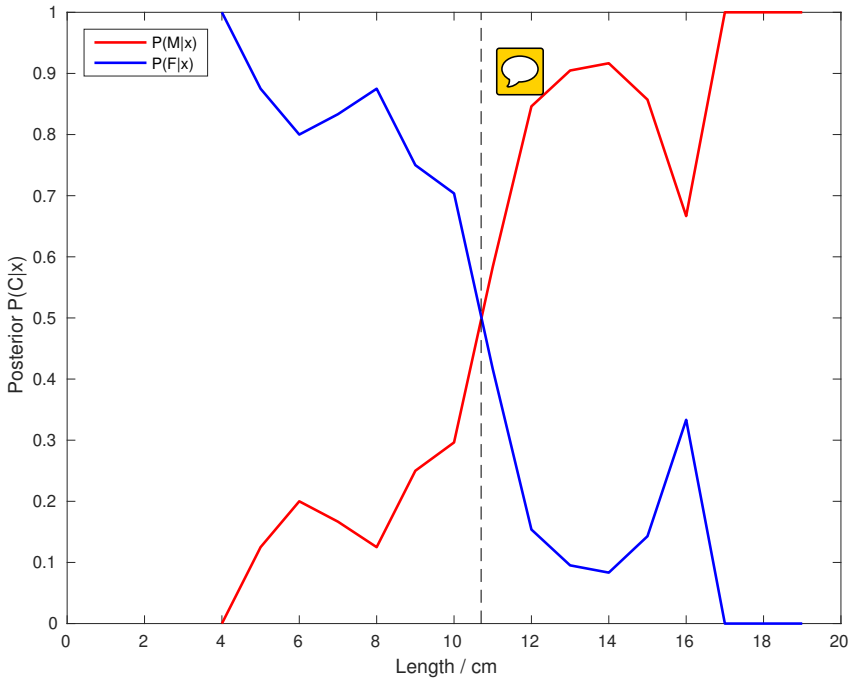

Figure 3: *Posterior probabilities for male and female fish lengths. The vertical black line around* $x = 11$ *denotes the decision boundary.*

Figure 3 depicts how the posterior probability changes as a function of the fish length for both the male (red) and female (blue) fish. The vertical black line denotes the decision boundary (i.e. where the posterior probabilities of male and female fish are the same). If we used these probabilities to assign a new (unlabelled) fish, we would classify it as `female` if its length was on the left of the decision boundary, and `male` otherwise.

Assume that we observe a new fish for which we know that has length $x = 8$. Should we classify it as male or female fish?

```
% Fish of length x = 8, are in the 5th element of the likelihood vectors
% So we compute the test point likelihood directly from that element
>> test_lik_M = lik_M(5);
>> test_lik_F = lik_F(5);
% Compute posterior probabilities for each class
>> test_post_M = test_lik_M * prior_M / Px(5)
test_post_M =
    0.1250
>> test_post_F = test_lik_F * prior_F / Px(5)
test_post_F =
    0.8750
```

Hence the fish would be classified as female, which could be observed directly from Figure 3.

1.4 Bayes decision rule

In the previous section (for the sake of illustration) we computed the actual posterior probabilities for each class and then assigned each example to the class with the maximum posterior probability. However, computing posterior probabilities for real life problems is often impractical, mainly due the denominator in the Bayes theorem (i.e. the evidence).

Since our goal is to classify a test example to the most probable class, we can compute their ratio:

$$\frac{P(C_M|x)}{P(C_F|x)} = \frac{\frac{P(x|C_M)P(C_M)}{P(x)}}{\frac{P(x|C_F)P(C_F)}{P(x)}} = \frac{P(x|C_M)P(C_M)}{P(x|C_F)P(C_F)} \tag{5}$$

If the ratio in the above equation is greater than 1 then x is classified as M, if x is less than 1 then x is classified as F. As you can observe, the denominator term $P(x)$ cancels, so there is no need to compute it at all.

Let's compute the ratio of the above example for a test fish of length $x = 8$. We would expect the ratio to be less than 1, since the fish should be classified as female.

```
% Compute ratio of posterior probabilities for test example x = 8
>> test_ratio = (test_lik_M * prior_M) / (test_lik_F * prior_F)
test_ratio =
    0.1429
```

Exercises

- Compute the posterior probabilities for each class using the following prior distributions $P(C_M) = 0.9$ and $P(C_F = 0.1)$. Create the likelihood and the posterior probability plots as shown in the previous sections. What do you observe? Does the likelihood depend on the prior?
- Classify the test example $x = 8$ using the updated posterior probabilities.
- Assuming equal prior probabilities, classify the following test examples: $x = 2, 9, 12, 16$.