Tutorial 3: Naive Bayes and Gaussians

1. The total number of documents in the training set is $N = 11$, with $N_S = 6$, $N_I = 5$.

   We can estimate the prior probabilities from the training data as:

$$P(S) = \frac{N_S}{N} = \frac{6}{11}; \qquad P(I) = \frac{N_I}{N} = \frac{5}{11}.$$

   Let $n_S(w)$ be the frequency of word w in all documents of class S, giving likelihood estimate without smoothing and with add-one smoothing:

$$\hat{P}(w|S) = \frac{n_S(w)}{\sum_{v \in V} n_S(v)},$$
$$\tilde{P}(w|S) = \frac{n_S(w) + 1}{|V| + \sum_{v \in V} n_S(v)},$$

   where V is the vocabulary (set of word types under consideration).

	$n_S(w)$	$\hat{P}(w\|S)$	$\tilde{P}(w\|S)$	$n_I(w)$	$\hat{P}(w\|I)$	$\tilde{P}(w\|I)$
w_1	6	6/36	7/44	1	1/16	2/24
w_2	0	0/36	1/44	4	4/16	5/24
w_3	2	2/36	3/44	3	3/16	4/24
w_4	5	5/36	6/44	1	1/16	2/24
w_5	4	4/36	5/44	1	1/16	2/24
w_6	6	6/36	7/44	2	2/16	3/24
w_7	7	7/36	8/44	3	3/16	4/24
w_8	6	6/36	7/44	1	1/16	2/24

   We have now estimated the model parameters.

   (a) $\mathcal{D}_1 = w_5\ w_1\ w_6\ w_8\ w_1\ w_2\ w_6$

Case-1 (no smoothing)

$$P(\mathcal{D}_1|S) = P(w_5|S) \cdot P(w_1|S) \cdot P(w_6|S) \cdot P(w_8|S) \cdot P(w_1|S) \cdot P(w_2|S) \cdot P(w_6|S)$$
$$= \frac{4}{36} \times \frac{6}{36} \times \frac{6}{36} \times \frac{6}{36} \times \frac{6}{36} \times \frac{0}{36} \times \frac{6}{36} = 0$$
$$P(S|\mathcal{D}_1) \propto P(S)P(\mathcal{D}_1|S) = 0$$
$$P(\mathcal{D}_1|I) = P(w_5|I) \cdot P(w_1|I) \cdot P(w_6|I) \cdot P(w_8|I) \cdot P(w_1|I) \cdot P(w_2|I) \cdot P(w_6|I)$$
$$= \frac{1}{16} \times \frac{1}{16} \times \frac{2}{16} \times \frac{1}{16} \times \frac{1}{16} \times \frac{4}{16} \times \frac{2}{16} = \frac{2^4}{16^7}$$
$$P(I|\mathcal{D}_1) = \frac{P(I)P(\mathcal{D}_1|I)}{P(S)P(\mathcal{D}_1|S) + P(I)P(\mathcal{D}_1|I)} = 1$$

$P(I|\mathcal{D}_1) < P(S|\mathcal{D}_1)$, thus we classify $\mathcal{D}_1$ as I.

Case-2 (with smoothing)

$$P(\mathcal{D}_1|S) = P(w_5|S) \cdot P(w_1|S) \cdot P(w_6|S) \cdot P(w_8|S) \cdot P(w_1|S) \cdot P(w_2|S) \cdot P(w_6|S)$$
$$= \frac{5}{44} \times \frac{7}{44} \times \frac{7}{44} \times \frac{7}{44} \times \frac{7}{44} \times \frac{1}{44} \times \frac{7}{44} = \frac{84035}{44^7} = 2.63 \times 10^{-7}$$
$$P(S|\mathcal{D}_1) \propto P(S)\,P(\mathcal{D}_1|S) = \frac{6}{11} \cdot \frac{84035}{44^7} = 1.44 \times 10^{-7}$$

$$P(\mathcal{D}_1|I) = P(w_5|I) \cdot P(w_1|I) \cdot P(w_6|I) \cdot P(w_8|I) \cdot P(w_1|I) \cdot P(w_2|I) \cdot P(w_6|I)$$
$$= \frac{1}{12} \times \frac{1}{12} \times \frac{1}{8} \times \frac{1}{12} \times \frac{1}{12} \times \frac{5}{24} \times \frac{1}{8} = \frac{5}{31850496} = 1.57 \times 10^{-7}$$
$$P(I|\mathcal{D}_1) \propto P(I)\,P(\mathcal{D}_1|I) = \frac{5}{11} \cdot \frac{5}{31850496} = 7.14 \times 10^{-8}$$

$P(S|\mathcal{D}_1) > P(I|\mathcal{D}_1)$, thus we classify $\mathcal{D}_1$ as S.

We have not normalised by $P(\mathcal{D}_1)$, hence the above are joint probabilities, proportional to the posterior probability. To obtain the posterior:

$$P(S|\mathcal{D}_1) = \frac{P(S)\,P(\mathcal{D}_1|S)}{P(S)\,P(\mathcal{D}_1|S) + P(I)\,P(\mathcal{D}_1|I)} = \frac{1.44 \times 10^{-7}}{1.44 \times 10^{-7} + 7.14 \times 10^{-8}} = 0.67$$
$$P(I|\mathcal{D}_1) = 1 - P(S|\mathcal{D}_1) = 0.33$$

(b) $\mathcal{D}_2 = w_3\ w_5\ w_2\ w_7$

Case-1 (no smoothing)

$$P(\mathcal{D}_2|S) = P(w_3|S) \cdot P(w_5|S) \cdot P(w_2|S) \cdot P(w_7|S)$$
$$= \frac{2}{36} \times \frac{4}{36} \times \frac{0}{36} \times \frac{7}{36} = 0$$
$$P(S|\mathcal{D}_2) \propto P(S)P(\mathcal{D}_2|S) = 0$$
$$P(\mathcal{D}_2|I) = P(w_3|I) \cdot P(w_5|I) \cdot P(w_2|I) \cdot P(w_7|I)$$
$$= \frac{3}{16} \times \frac{1}{16} \times \frac{4}{16} \times \frac{3}{16} = \frac{2^2 \times 3^2}{16^4}$$
$$P(I|\mathcal{D}_2) = \frac{P(I)P(\mathcal{D}_2|I)}{P(S)P(\mathcal{D}_2|S) + P(I)P(\mathcal{D}_2|I)} = 1$$

Case-2 (with smoothing)

$$P(\mathcal{D}_2|S) = P(w_3|S) \cdot P(w_5|S) \cdot P(w_2|S) \cdot P(w_7|S)$$
$$= \frac{3}{44} \times \frac{5}{44} \times \frac{1}{44} \times \frac{2}{11} = \frac{30}{937024} = 3.20 \times 10^{-5}$$
$$P(S|\mathcal{D}_2) \propto P(S)\,P(\mathcal{D}_2|S) = \frac{6}{11} \cdot \frac{30}{937024} = 1.75 \times 10^{-5}$$
$$P(\mathcal{D}_2|I) = P(w_3|I) \cdot P(w_5|I) \cdot P(w_2|I) \cdot P(w_7|I)$$
$$= \frac{1}{6} \times \frac{1}{12} \times \frac{5}{24} \times \frac{1}{6} = \frac{5}{10368} = 4.82 \times 10^{-4}$$
$$P(I|\mathcal{D}_2) \propto P(I)\,P(\mathcal{D}_2|I) = \frac{5}{11} \cdot \frac{5}{10368} = 2.19 \times 10^{-4}$$

$P(I|\mathcal{D}_2) > P(S|\mathcal{D}_2)$, thus we classify $\mathcal{D}_1$ as I.

We have not normalised by $P(\mathcal{D}_2)$, hence the above are joint probabilities, proportional to the posterior probability. To obtain the posterior:

$$P(S|\mathcal{D}_2) = \frac{P(S)\,P(\mathcal{D}_2|S)}{P(S)\,P(\mathcal{D}_2|S) + P(I)\,P(\mathcal{D}_2|I)}$$
$$= \frac{1.75 \times 10^{-5}}{1.75 \times 10^{-5} + 2.19 \times 10^{-4}} = 0.074$$
$$P(I|\mathcal{D}_2) = 1 - P(S|\mathcal{D}_2) = 0.926$$

2. (a) Since rand() generates samples based on a uniform distribution on the interval $(0, 1)$, we can assign samples on $(0, 0.5)$ to 'H' and, those on $[0.5, 1)$ to 'T' to simulate tossing a fair coin, which is illustrated in Figure 2 (a).

   Here is a sample Matlab sample:

```
N = 10;          % the number of trials
threshold = 0.5;
for i = 1 : N
   v = rand();
   if(v < threshold)
      x = 'H';
   else
      x = 'T';
   end
   fprintf(1, '%s ', x);
end
fprintf(1, '\n');
```

   (b) We now need to change the widths of the intervals so that they correspond to $P(H) = 0.6$ and $P(T) = 0.4$, which are shown in Figure 2 (b). The Matlab code is basically the same as the above, except 'threshold = 0.6' now.

   (c) We split the interval $(0, 1)$ into six disjoint regions, each of which corresponds to $P(i), i = 1, \ldots, 6$. Here is a sample Matlab code:

```
N = 10;          % the number of trials
P = [0.05, 0.1, 0.14, 0.19, 0.24, 0.28];   % P(X)
n_faces = length(P);
thresholds = cumsum(P) / sum(P);   % get cumulative sum of P()
for i = 1 : N
   v = rand();
   x = n_faces;
   for k = 1 : n_faces-1
      if( v < thresholds(k) )
         x = k;
         break;
      end
   end
   fprintf(1, '%d ', x);
end
fprintf(1, '\n');
```

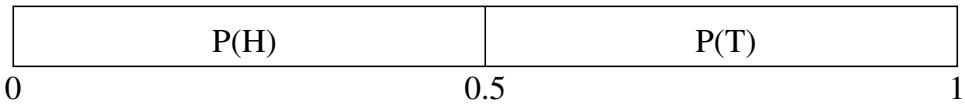


Figure 2 (a)

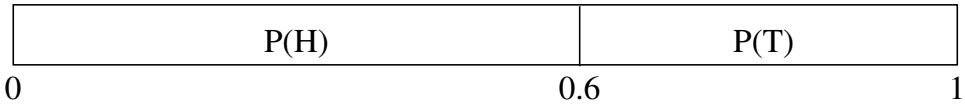


Figure 2 (b)

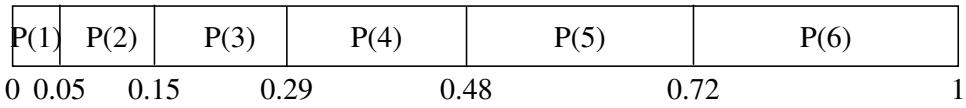
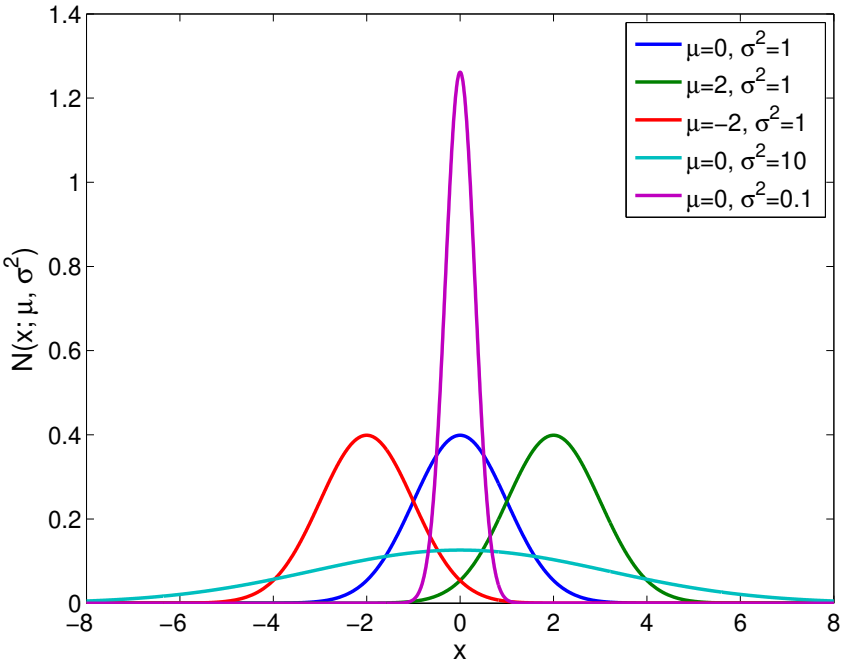


Figure 2 (c)

3.

(a) The sketch will look like this:



(b) As the pdf of a normal distribution is given by

$$p(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

it is easy to see that the width of the curve scales linearly with σ (not σ^2), and the height of the peak is proportional to the reciprocal of σ. Note that the exact height is $1/\left(\sigma\sqrt{2\pi}\right)$, which can be greater than 1 for small σ. See the figure above:

(c) Here is a sample Matlab code:

```
% Parameters of normal distributions to plot
% Each line represents the two paramters (mean, variance)
params = [
            0.0, 1.0;
            2.0, 1.0;
            -2.0, 1.0;
            0.0, 10.0;
            0.0, 0.1;
];

xrange = [-8, 8];        % x-range
np = 200;                % plotting resolution, i.e. number of points

x = linspace(xrange(1), xrange(2), np);
n_distributions = size(params,1);
X = zeros(n_distributions, length(x));
Y = X;
ss = cell(n_distributions, 1);
for i = 1 : n_distributions
  m = params(i,1); var = params(i,2);
  Y(i,:) = 1/(sqrt(2*pi*var)) * exp(-(x-m).^2 ./ (2*var));
  X(i,:) = x;
  ss{i} = sprintf('\\mu=%g, \\sigma^2=%g', m, var);
end

plot(X', Y', 'linewidth', 2);
set(gca, 'fontsize', 14);
xlabel('x', 'fontsize', 16);
ylabel('N(x; \mu, \sigma^2)', 'fontsize', 16);
legend(ss, 'fontsize', 14);
```

4. (a)

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n, \qquad \hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \hat{\boldsymbol{\mu}})(\mathbf{x}_n - \hat{\boldsymbol{\mu}})^T$$

$$\hat{\boldsymbol{\mu}} = \frac{1}{4}\left\{\begin{pmatrix}1\\1\end{pmatrix}+\begin{pmatrix}1\\2\end{pmatrix}+\begin{pmatrix}3\\2\end{pmatrix}+\begin{pmatrix}3\\3\end{pmatrix}\right\} = \begin{pmatrix}2\\2\end{pmatrix}.$$

Subtracting the mean from each samples gives:

$$(-1,-1)^T; \quad (-1,0)^T; \quad (1,0)^T; \quad (1,1)^T$$

$$\begin{aligned}\hat{\boldsymbol{\Sigma}} &= \frac{1}{4}\left\{\begin{pmatrix}-1\\-1\end{pmatrix}\begin{pmatrix}-1, & -1\end{pmatrix}+\begin{pmatrix}-1\\0\end{pmatrix}\begin{pmatrix}-1, & 0\end{pmatrix}+\begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}1, & 0\end{pmatrix}+\begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}1, & 1\end{pmatrix}\right\}\\ &= \frac{1}{4}\left\{\begin{pmatrix}1&1\\1&1\end{pmatrix}+\begin{pmatrix}1&0\\0&0\end{pmatrix}+\begin{pmatrix}1&0\\0&0\end{pmatrix}+\begin{pmatrix}1&1\\1&1\end{pmatrix}\right\}\\ &= \frac{1}{4}\begin{pmatrix}4&2\\2&2\end{pmatrix} = \begin{pmatrix}1&1/2\\1/2&1/2\end{pmatrix}.\end{aligned}$$

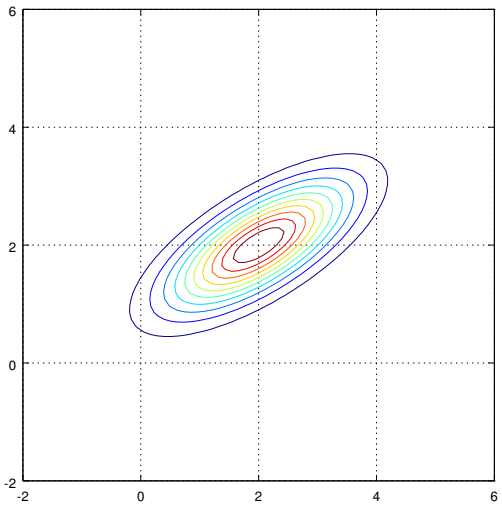
(b) Let $R = (\rho_{ij})$ to denote the correlation matrix.

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}\,.$$

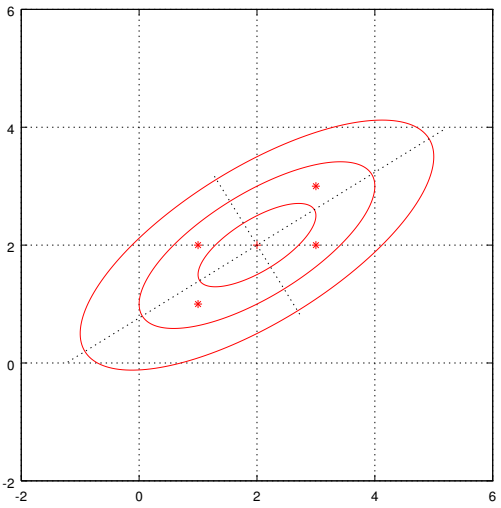
We will get $\rho_{11} = \rho_{22} = 1$ and $\rho_{12} = \rho_{21} = \frac{1/2}{\sqrt{1/2}} = \frac{\sqrt{2}}{2}$. Thus

$$R = \begin{pmatrix} 1 & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & 1 \end{pmatrix}.$$

(c) Examples of coutours are shown below, where (a) shows the contours drwan with Matlab's contour(), (b) shows the contours for standard deviations, 1, 2, and 3, and the two principal components, whose slopes are 0.618 and -1.618, respectively.



(a)



(b)

(d) The log likelihood of D-dimensional Gaussian distribution is given as

$$\ln p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\,.$$

In the present case, $D = 2$, $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$,

$$|\boldsymbol{\Sigma}| = \frac{1}{4},$$
$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} 2 & -2 \\ -2 & 4 \end{pmatrix}.$$

For $\mathbf{z} = (2,1)^T$, the 3rd term of the above log likelihood is gieven as follows.

$$-\frac{1}{2}\left(\begin{pmatrix} 2 \\ 1 \end{pmatrix} - \begin{pmatrix} 2 \\ 2 \end{pmatrix}\right)^T \begin{pmatrix} 2 & -2 \\ -2 & 4 \end{pmatrix}\left(\begin{pmatrix} 2 \\ 1 \end{pmatrix} - \begin{pmatrix} 2 \\ 2 \end{pmatrix}\right)$$
$$= -\frac{1}{2}\begin{pmatrix} 0 \\ -1 \end{pmatrix}^T \begin{pmatrix} 2 & -2 \\ -2 & 4 \end{pmatrix}\begin{pmatrix} 0 \\ -1 \end{pmatrix} = -\frac{1}{2}\begin{pmatrix} 0 & -1 \end{pmatrix}\begin{pmatrix} 2 \\ -4 \end{pmatrix} = -\frac{1}{2}\times 4 = -2\,.$$

Thus,

$$\ln p(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\ln(2\pi) + \ln 2 - 2 = -\ln(\pi) - 2 \approx -3.1447\,.$$

(e) We can still estimate the mean vector, covariance matrix, and correlation matrix, but the deteminant of the covariance matrix is zero. As a result, the inverse of the covariance matrix does not exist, and we are unable to apply the Gaussian pdf for 2D to find the log likeloood of $\mathbf{z}$. (The zero determinat means rank($\mathbf{\Sigma}$) < 2, which can be confirmed by the fact that the two samples lie in a single line in 2D. You can generalise this to D-dimensional case, so that $|\mathbf{\Sigma}| = 0$ for $N \leq D$, where N denotes the number of samples.)

5. First, we show that the mean is calculated correctly, where m_n is the mean of the first n values, and r_n is defined as

$$r_n = x_n - m_{n-1} \tag{1}$$

$$\begin{aligned} m_{n-1} &= \frac{1}{n-1}\sum_{i=1}^{n-1} x_i \\ m_n &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{1}{n}\sum_{i=1}^{n-1} x_i + \frac{x_n}{n} \\ &= \frac{n-1}{n} m_{n-1} + \frac{x_n}{n} \\ &= m_{n-1} - \frac{1}{n} m_{n-1} + \frac{x_n}{n} \\ &= m_{n-1} + \frac{x_n - m_{n-1}}{n} \\ &= m_{n-1} + \frac{r_n}{n}\,. \end{aligned} \tag{2}$$

Now for variance; define n times the variance as $S = \sum_{i=1}^{n} (x_i - m)^2$.

As before, taking m_n to be mean of first n values. Defining S_n to be n times the variance for first n values, that is:

$$\begin{aligned} S_n &= \sum_{i=1}^{n} (x_i - m_n)^2 \\ &= \sum_{i=1}^{n} ((x_i - m_{n-1}) + (m_{n-1} - m_n))^2 \\ &= \sum_{i=1}^{n} (x_i - m_{n-1})^2 + \sum_{i=1}^{n} (m_{n-1} - m_n)^2 + 2\sum_{i=1}^{n} (x_i - m_{n-1})(m_{n-1} - m_n)\,. \end{aligned} \tag{3}$$

Taking each of the three terms on the RHS in turn. The first term may be written, using (1):

$$\begin{aligned} \sum_{i=1}^{n} (x_i - m_{n-1})^2 &= \sum_{i=1}^{n-1} (x_i - m_{n-1})^2 + (x_n - m_{n-1})^2 \\ &= S_{n-1} + (x_n - m_{n-1})^2 \\ &= S_{n-1} + r_n^2\,. \end{aligned} \tag{4}$$

From (2) we can write:

$$m_n - m_{n-1} = \frac{r_n}{n}\,. \tag{5}$$

We can use this to rewrite the second term:

$$\begin{aligned}\sum_{i=1}^{n}(m_{n-1} - m_n)^2 &= n(m_{n-1} - m_n)^2 \\ &= \frac{r_n^2}{n}\,.\end{aligned} \tag{6}$$

And the third term, again using (5):

$$\begin{aligned}2\sum_{i=1}^{n}(x_i - m_{n-1})(m_{n-1} - m_n) &= 2(m_{n-1} - m_n)\sum_{i=1}^{n}(x_i - m_{n-1}) \\ &= \frac{-2r_n}{n}\sum_{i=1}^{n}(x_i - m_{n-1}) \\ &= \frac{-2r_n}{n}\left(\sum_{i=1}^{n}x_i - nm_{n-1}\right) \\ &= \frac{-2r_n}{n}(nm_n - nm_{n-1}) \\ &= \frac{-2r_n^2}{n}\,.\end{aligned} \tag{7}$$

Substituting (4), (6) and 7 into (3):

$$\begin{aligned}S_n &= S_{n-1} + r_n^2 + \frac{r_n^2}{n} - \frac{2r_n^2}{n} \\ &= S_{n-1} + \frac{(n-1)}{n}r_n^2 \\ &= S_{n-1} + \left(1 - \frac{1}{n}\right)r_n^2\,.\end{aligned} \tag{8}$$