Tutorial 6: 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 $n(w, S)$ be the frequency of word w in all documents of class S, giving likelihood estimate,

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

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

	$n(w,S)$	$\hat{P}(w\|S)$	$n(w,I)$	$\hat{P}(w\|I)$
w_1	6	7/44	1	1/12
w_2	0	1/44	4	5/24
w_3	2	3/44	3	1/6
w_4	5	3/22	1	1/12
w_5	4	5/44	1	1/12
w_6	6	7/44	2	1/8
w_7	7	2/11	3	1/6
w_8	6	7/44	1	1/12

   We have now estimated the model parameters.

   (a) $D_1 = w_5\ w_1\ w_6\ w_8\ w_1\ w_2\ w_6$

$$\begin{aligned}
P(D_1|S) &= P(w_5|S) \cdot P(w_1|S) \cdot P(w_6|S) \cdot P(w_8|S) \cdot P(w_1|S) \cdot P(w_2|S) \cdot P(w_6|S) \\
&= \frac{5}{44} \times \frac{7}{44} \times \frac{7}{44} \times \frac{7}{44} \times \frac{7}{44} \times \frac{1}{44} \times \frac{7}{44} \\
&= \frac{84035}{44^7} = 2.63 \times 10^{-7} \\
P(S|D_1) &\propto P(S)\, P(D_1|S) \\
&= \frac{6}{11} \cdot \frac{84035}{44^7} = 1.44 \times 10^{-7}
\end{aligned}$$

$$\begin{aligned}
P(D_1|I) &= P(w_5|I) \cdot P(w_1|I) \cdot P(w_6|I) \cdot P(w_8|I) \cdot P(w_1|I) \cdot P(w_2|I) \cdot P(w_6|I) \\
&= \frac{1}{12} \times \frac{1}{12} \times \frac{1}{8} \times \frac{1}{12} \times \frac{1}{12} \times \frac{5}{24} \times \frac{1}{8} \\
&= \frac{5}{31850496} = 1.57 \times 10^{-7} \\
P(I|D_1) &\propto P(I)\, P(D_1|I) \\
&= \frac{5}{11} \cdot \frac{5}{31850496} = 7.14 \times 10^{-8}
\end{aligned}$$

   $P(S|D_1) > P(I|D_1)$, thus we classify D_1 as S.

   We have not normalised by $P(D_1)$, hence the above are joint probabilities, proportional to the posterior probability. To obtain the posterior:

$$\begin{aligned}
P(S|D_1) &= \frac{P(S)\, P(D_1|S)}{P(S)\, P(D_1|S) + P(I)\, P(D_1|I)} \\
&= \frac{1.44 \times 10^{-7}}{1.44 \times 10^{-7} + 7.14 \times 10^{-8}} = 0.67 \\
P(I|D_1) &= 1 - P(S|D_1) = 0.33
\end{aligned}$$

   (b) $D_2 = w_3\ w_5\ w_2\ w_7$

$$\begin{aligned}
P(D_2|S) &= P(w_3|S) \cdot P(w_5|S) \cdot P(w_2|S) \cdot P(w_7|S) \\
&= \frac{3}{44} \times \frac{5}{44} \times \frac{1}{44} \times \frac{2}{11} \\
&= \frac{30}{937024} = 3.20 \times 10^{-5} \\
P(S|D_2) &\propto P(S)\, P(D_2|S) \\
&= \frac{6}{11} \cdot \frac{30}{937024} = 1.75 \times 10^{-5}
\end{aligned}$$

$$\begin{aligned}
P(D_2|I) &= P(w_3|I) \cdot P(w_5|I) \cdot P(w_2|I) \cdot P(w_7|I) \\
&= \frac{1}{6} \times \frac{1}{12} \times \frac{5}{24} \times \frac{1}{6} \\
&= \frac{5}{10368} = 4.82 \times 10^{-4} \\
P(I|D_2) &\propto P(I)\, P(D_2|I) \\
&= \frac{5}{11} \cdot \frac{5}{10368} = 2.19 \times 10^{-4}
\end{aligned}$$

   $P(I|D_2) > P(S|D_2)$, thus we classify D_1 as I.

   We have not normalised by $P(D_2)$, hence the above are joint probabilities, proportional to the posterior probability. To obtain the posterior:

$$\begin{aligned}
P(S|D_2) &= \frac{P(S)\, P(D_2|S)}{P(S)\, P(D_2|S) + P(I)\, P(D_2|I)} \\
&= \frac{1.75 \times 10^{-5}}{1.75 \times 10^{-5} + 2.19 \times 10^{-4}} = 0.074 \\
P(I|D_2) &= 1 - P(S|D_2) = 0.926
\end{aligned}$$

How would the classifications differ if add-one smoothing had not been used when estimating the model parameters?

Since $n(w_2, S) = 0$, if smoothing was not used, then $\hat{P}(w_2 \mid S)$ would have been estimated as 0. In which case, since w_2 occurs in both test documents, both $P(D_1 \mid S)$ and $P(D_2 \mid S)$ would have been estimated as 0, and hence $P(S \mid D_1)$ and $P(S \mid D_2)$ would both have been computed as 0, so both documents would have been classified as I (with a posterior probability of 1).

2. Let x be a word type with count 10, y be a word type with count 5, and z be a word type with count 0.

   (a) 12 word vocab:

$$P_{RF}(x) = \frac{10}{100} = 0.1 \qquad P_{Lap}(x) = \frac{11}{112} = 0.098 \qquad P_{AD}(x) = \frac{9.7}{100} = 0.097$$

$$P_{RF}(y) = \frac{5}{100} = 0.05 \qquad P_{Lap}(y) = \frac{6}{112} = 0.054 \qquad P_{AD}(y) = \frac{4.7}{100} = 0.047$$

$$P_{RF}(z) = \frac{0}{100} = 0 \qquad P_{Lap}(z) = \frac{1}{112} = 0.0089 \qquad P_{AD}(z) = \frac{0.3 \cdot 11/1}{100} = 0.033$$

   (b) 20 word vocab:

$$P_{RF}(x) = \frac{10}{100} = 0.1 \qquad P_{Lap}(x) = \frac{11}{120} = 0.092 \qquad P_{AD}(x) = \frac{9.7}{100} = 0.097$$

$$P_{RF}(y) = \frac{5}{100} = 0.05 \qquad P_{Lap}(y) = \frac{6}{120} = 0.05 \qquad P_{AD}(y) = \frac{4.7}{100} = 0.047$$

$$P_{RF}(z) = \frac{0}{100} = 0 \qquad P_{Lap}(z) = \frac{1}{120} = 0.0083 \qquad P_{AD}(z) = \frac{0.3 \cdot 11/9}{100} = 0.0037$$

   (c) 1000 word vocab:

$$P_{RF}(x) = \frac{10}{100} = 0.1 \qquad P_{Lap}(x) = \frac{11}{1100} = 0.01 \qquad P_{AD}(x) = \frac{9.7}{100} = 0.097$$

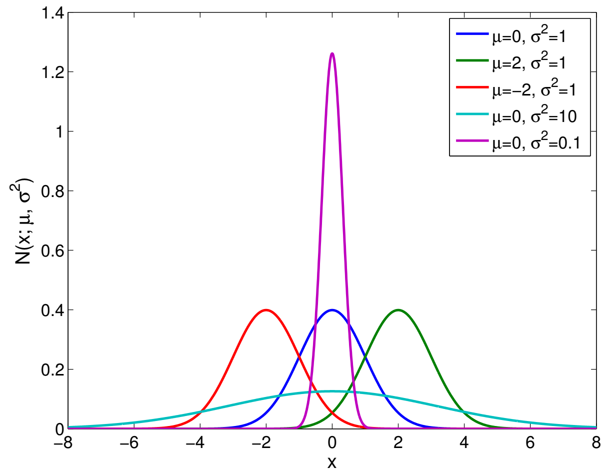
$$P_{RF}(y) = \frac{5}{100} = 0.05 \qquad P_{Lap}(y) = \frac{6}{1100} = 0.0055 \qquad P_{AD}(y) = \frac{4.7}{100} = 0.047$$

$$P_{RF}(z) = \frac{0}{100} = 0 \qquad P_{Lap}(z) = \frac{1}{1100} = 0.00091 \qquad P_{AD}(z) = \frac{0.3 \cdot 11/989}{100} = 0.000033$$

Key points to note:

- For the count 10 items, the add-one probability estimate (from the same sample) decreases by a factor of 10 when the number of unknown items is increased from 1 to 989!
- On the other hand, the estimate for the observed items is stable with absolute discounting (but the probability for unobserved items get smeared thinly across however many there are).
- Add one smoothing tends to overestimate the probabilities of unseen events. In this example in part (a) 0.0089 is allocated to unseen events (there is only one); in part (b) where there are 9 unknown word types, $9 \times 0.0083 = 0.0747$ is allocated to unseen events — this already rather high (7.5%!); in part (c) where there are 989 unknown word types, $989 \times 0.00091 = 0.90$ is allocated to unseen events — 90% of the probability is allocated to unseen events, whereas only 10% is used for observed events!
- In general, when the number of samples is much greater than the number of events, add one smoothing can be OK. Otherwise it can grossly over-estimate the probability for unseen events.
- (In language modelling for speech recognition or machine translation, where we we estimate probabilities of triples of words, the number of events might be $50\,000^3 \sim 10^{14}$).
- Absolute discounting works much better since it does not add counts, it just reallocates the existing counts.
- (More sophisticated versions of absolute discounting estimate k using the number of events with counts 1 or 2 — the intuition is that the best way to estimate the probability of events that have not occurred is to look at observed but very infrequent events — e.g.: $k \sim u(1)/(u(1) + 2u(2))$).

3. (a) The sketch will look like this:



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

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

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. (The exact height is $1/(\sigma\sqrt{2\pi})$. Note that the height can be greater than 1. 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. 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:

$$S_n = \sum_{i=1}^{n}(x_i - m_n)^2$$

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

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):

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

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

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