**Inf2b Learning and Data**

*Lecture: Revision meeting*

Hiroshi Shimodaira

(Credit: Iain Murray and Steve Renals)

Centre for Speech Technology Research (CSTR)
School of Informatics
University of Edinburgh

Jan-Mar 2015

---

**Revision**

- Past exam papers
- Examples in notes, slides, and tutorials
- Review definitions, formulae and algorithms to understand the concepts. Just remembering solutions and techniques is no good.

---

**Solutions for past exam papers**

<table>
<thead>
<tr>
<th>April</th>
<th>August</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013/14</td>
<td>Yes</td>
</tr>
<tr>
<td>2012/13</td>
<td>-</td>
</tr>
<tr>
<td>2011/12</td>
<td>-</td>
</tr>
<tr>
<td>2010/11</td>
<td>-</td>
</tr>
<tr>
<td>2009/10</td>
<td>Yes</td>
</tr>
<tr>
<td>2008/09</td>
<td>Yes</td>
</tr>
<tr>
<td>2007/08</td>
<td>-</td>
</tr>
<tr>
<td>2006/07</td>
<td>Yes</td>
</tr>
<tr>
<td>2005/06</td>
<td>-</td>
</tr>
</tbody>
</table>

---

**Question 1**

*Can we rate systems (movies) based on normalized scores \( z_{cm} \) (\( c \) being rating given by critic \( c \) to movie \( m \)) by comparing said scores using Euclidean distances to that of the user’s scores? If not, why exactly?*

- What do you think, at first?
- Yes, you can, it would work better than just using the original scores.
- However, if users scores are not normalised, there will be a potential mismatch in getting Euclidean distance between a critic and the user, about which you need to consider.

**Question 2**

*How do we use PCA to reduce dimensionality? Moreover, how do we do that without losing information in the feature vector (keeping a set of them distinct even after dimensionality reduction)?*

\[
\begin{bmatrix}
\mathbf{y}_1 \\
\vdots \\
\mathbf{y}_n
\end{bmatrix} = \begin{bmatrix}
\mathbf{p}_{1}^T \\
\vdots \\
\mathbf{p}_{d}^T
\end{bmatrix} \begin{bmatrix}
\mathbf{x}
\end{bmatrix} = \mathbf{A}^T \mathbf{x}
\]

where \( \ell \leq d \)

- The information will be lost if \( \ell < d \).
- For details, see Lecture-3 slides

---

**Question 3**

*How much do we need to know about eigenvectors?*

You are supposed to know how they are used for PCA

---

**Question 4**

*When using K-NN, if there is a tie between classes, do we choose the resulting class randomly, or same as in the coursework (the class whose representative is closest to the item, whose class is in question)?*

Either is fine.

---

**Question 5**

*In 2013 April exam, We were asked to “Describe a way to pre-process the data to make the algorithm [K-means clustering] invariant to the choice of linear scale used to measure each feature” in a graph. I believe the solution lies in the Euclidean distances for K-means clustering. Do I perhaps apply PCA to the problem to reduce the dimensionality and focus only on axes with the same scale (eliminating ones with different scales)?*

- You are close to the right answer.
- See page 24 of Lecture-3 slides, and consider a normalisation technique which yields scale invariance.
**Question 6**

Do we need to apply +1 smoothing for the Bernoulli document model and if yes, then how?

- Not really, but you can. (cf. Multinomial doc. model.)
- You will need to consider smoothing

**Priors:**  \( P(c_k) \approx \frac{N_k}{N} \)

**Likelihoods:**  \( P(w_t | c_k) \approx \frac{n_k(w_t)}{N_k} \) (fraction of class k docs with word \( w_t, t = 1, \ldots, d \))

**Question 6 (cont.)**

Can you over the zero probability problem for the multinomial document model - especially what to add in the denominator and why we add that. (I have seen different versions online - for instance when we add 2 in the denominator).

- Zero probability problem: See Lecture-7 slides
- Understand the impact of zero probability problem for multinomial document model and Bernoulli document model

**Question 6 (cont.)**

- The purposes of smoothing
  - Avoid zero probability problem (serious in multinomial model)
  - Better estimate true probabilities (priors and likelihoods) from samples (observations).
- There are various smoothing methods, but none of them is perfect. (we don’t know the true distributions)

**Question 7**

Can you go over the formula for how to calculate the posterior probability of a document belonging to a class?

- For Bernoulli document model: See pp.9–13 of Lecture-7 slides
- For multinomial document model: See pp.16–17 of Lecture-7 slides and Q1 of Tutorial 6

**Question 8**

How is the 2-class discriminant and the linear one related?

Linear discriminant functions should be considered as a subset of discriminant functions.

- discriminant function: \( g_c(x) \)
- linear discriminant function: \( w_k^T x + w_0 = w_k x_1 + \cdots + w_k x_d + w_0 \)
- 2-class discriminant function: \( g(x) \)

**Question 9**

How do we plot Gaussian distributions if given a Mean and Standard Deviation?

\[
N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)
\]

Matlab/Octave code:

```matlab
mu = 0.5; % mean
sigma = 1.0; % sigma (i.e. standard deviation)
x = linspace(-4, 4, 100);
y = 1/(sqrt(2*pi)*sigma) * exp(-(x-mu).^2/(2*sigma^2));
plot(x, y);
```

**Question 10**

When can we use the spherical Gaussian formulas?

Assuming you are asking practical situations. You can use spherical Gaussian.

- if the model reasonably fits the data (or if it is a reasonable approximation of a (true) covariance matrix)
- if the number of training data is (extremely) insufficient to estimate a full (or even an diagonal) covariance matrix, and you think variances are similar to each other, and you think simplicity (or computation speed/memory) is the primal factor rather than accuracy.
- You should also recall it is related to template-based classification using Euclidean distance (see p.22 of Lecture-10 slides), which is a very popular method that you might want to try first before trying more complex ones.

**Question 11**

I failed to see what impact Naive Bayes would have on Gaussian covariance matrices. From memory I wanted to say that covariance matrices for all classes in a Gaussian model would be the same, but I could not convince myself this.

- Recall the Naive Bayes rule/model first.
  - \( x_1, \ldots, x_d \) are independent from each other.
  - \( P(x_1, x_2, \ldots, x_d | c_k) \approx P(x_1 | c_k) \cdots P(x_d | c_k) \)
- See Lecture-11 slides (e.g. page 18) and Lecture-5 slides (page 28)
- See Lecture-8 slides (pp.26-32)
- NB: sharing covariance is a technique to reduce the number of parameters to estimate (by assuming covariance matrices are class-independent)

**Question 12**

How exactly to calculate the class independent covariance matrix (when each class shares the same full covariance matrix) On the coursework, I first found the unique covariance matrix for each class, and then found their average to use as the class independent covariance matrix - is this a correct way to do it?

- See Q2 in the coursework FAQ web page.
- See Lecture-11 slides (page 17).
- Special care would be needed if class sizes are different from each other.
Question 13
If you could go over question 5, part b (about Gaussians) from the May 2011 past paper. “Show that the maximum likelihood estimate for the variance is indeed the sample variance.”
- See pp.20–22 of Lecture-8 slides
- NB: “sample variance” in the exam question indicates what is usually referred to as ‘population variance’ - the one which employs $N$ as the normaliser rather than $N - 1$.

Question 14
If you can go over two-class linear discriminants and the basic idea of neural networks in the revision lecture tomorrow, that would be great.
For neural networks,
- See Lectures 12 – 15 slides
- http://neuralnetworksanddeeplearning.com/

Question 15
Are multiple layer neural networks examinable?
Yes, everything in the lecture notes and slides are examinable unless otherwise noted.

Question 16
What is linear regression and generative modelling
- Linear regression: See Lecture-12 slides
- Generative models: See Lecture-7 slides