COMP229: Introduction to Data Science
Lecture 30: revision and exam advice

Vitaliy Kurlin, vitaliy.kurlin@liverpool.ac.uk
Autumn 2018, Computer Science department
University of Liverpool, United Kingdom

What hardest topics did you learn?

Short revisions were at the end of every lecture.

Two methods to reduce the dimension of data:

- Linear regression from Statistics: Lecture 15 discussed only dimension 2, a similar approach extends to higher dimensions.
- PCA (Principal Component Analysis, also SVD) based on Linear Algebra: Lecture 23.

How different are these methods, e.g. do they produce different results for data samples in $\mathbb{R}^2$?

The data example from Lecture 23

Lecture 23 has found the first principal direction $\vec{v}_1 = (\sqrt{5}+1, 2, 0)$ for the 5-point cloud in $\mathbb{R}^3$.

Subjects	student 1	student 2	3	4	5	mean
Maths x	5	5	4	3	3	$\bar{x} = 4$
English y	4	5	4	4	3	$\bar{y} = 4$

Let's forget about the 3rd coordinate and find the linear regression line $y = ax + b$ for the 5 points projected to $\mathbb{R}^2$: Maths (x) and English (y).

What are the formulae for the coefficients a and b?

Revision of the linear regression

Definition 15.1. For a scatterplot of n data points (x_i, y_i), the *least-squares regression line* has an equation $y = ax + b$ that minimises the sum of squares $f(a, b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2$.

Claim 15.2. The coefficients are $a = r_{xy}\frac{s_y}{s_x}$ and $b = \bar{y} - a\bar{x}$, where $\bar{x}$, $\bar{y}$ are the sample means and $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$, $s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$, the sample correlation $r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$.

Linear regression vs PCA

If you subtract the means, further computations will be easier: $a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x^2} = \frac{1 \cdot 0 + 1 \cdot 1 + 0 \cdot 0 + (-1) \cdot 0 + (-1) \cdot (-1)}{4} = 0.5.$

$b = 4 - 0.5 \cdot 4 = 2$, so the regression line is $y = 0.5x + 2$ and differs from the first principal direction parallel to $(\sqrt{5} + 1, 2)$. PCA minimises the variance (sum of squared orthogonal distances to the final line), but the linear regression $y(x)$ minimises the sum of squared vertical distances.

Advice already given in Lecture 1

Non-native English speakers can spend at least 1 hour per day on improving English skills: online, on paper, talking to people in university clubs, ...

Edit the COMP229 wiki page, e.g. by giving links to external resources. Your excellent contributions can be highlighted in references. Your career will depend on your connections, not on exam marks.

Your mark for the module = 100% exam consisting of 50% multiple choice, 50% written questions.

Time: 2.5 hours, all questions will be marked.

Advice on revisions before the exam

If you can suggest better hints, feel free to e-mail or share your advice on the discussion board.

- Short regular revisions are better (1-2 hours per module each working day) than a non-stop rush few days before the exam.
- Complement your brain activity by regular and substantial physical exercises.
- Sleep after lectures, not during lectures, because your brain during a good sleep strengthens all new neural connections.

What to expect at the exam

20 MCQ questions equally contribute to 50% of 100 marks. What's the mark per question?

5 written questions equally contribute to 50% of 100 marks. What's the mark per question?

One of the written questions asks for a proof.

All written questions ask for some definitions.

Calculators and tables of normal distributions aren't allowed to avoid the waste of your time.

Instructions on the exam booklet

INSTRUCTIONS TO CANDIDATES

HANDWRITING: If the examiner cannot read your handwriting you may be awarded a mark of zero.

You must not commence writing until instructed to do so by an invigilator.

1. Fill in the above lines before beginning to answer the questions.
2. Where it is practicable to do so, write on BOTH sides of the page. Where rough work is necessary, use the left-hand page for this purpose and cross out before handing in the book. *No part of this book may be torn off.*
3. Begin each answer on a new page and write the number of the question in the margin. If an answer requires more than one page, show the number of the question on each page.
4. If two or more books are used, fasten them together. Fasteners may be obtained from the Invigilators.
5. Fold over and seal the gummed edges to conceal your name.
6. If during the exam you have any questions or feel ill please inform the invigilator.

No books or papers may be taken to an examination desk. Your mobile phone and any other electronic data storage devices (including smartwatches) must be switched off and placed in a clear bag under your seat for the duration of the exam. Examination writing books, whether used or unused, must not be taken out of the examination room.

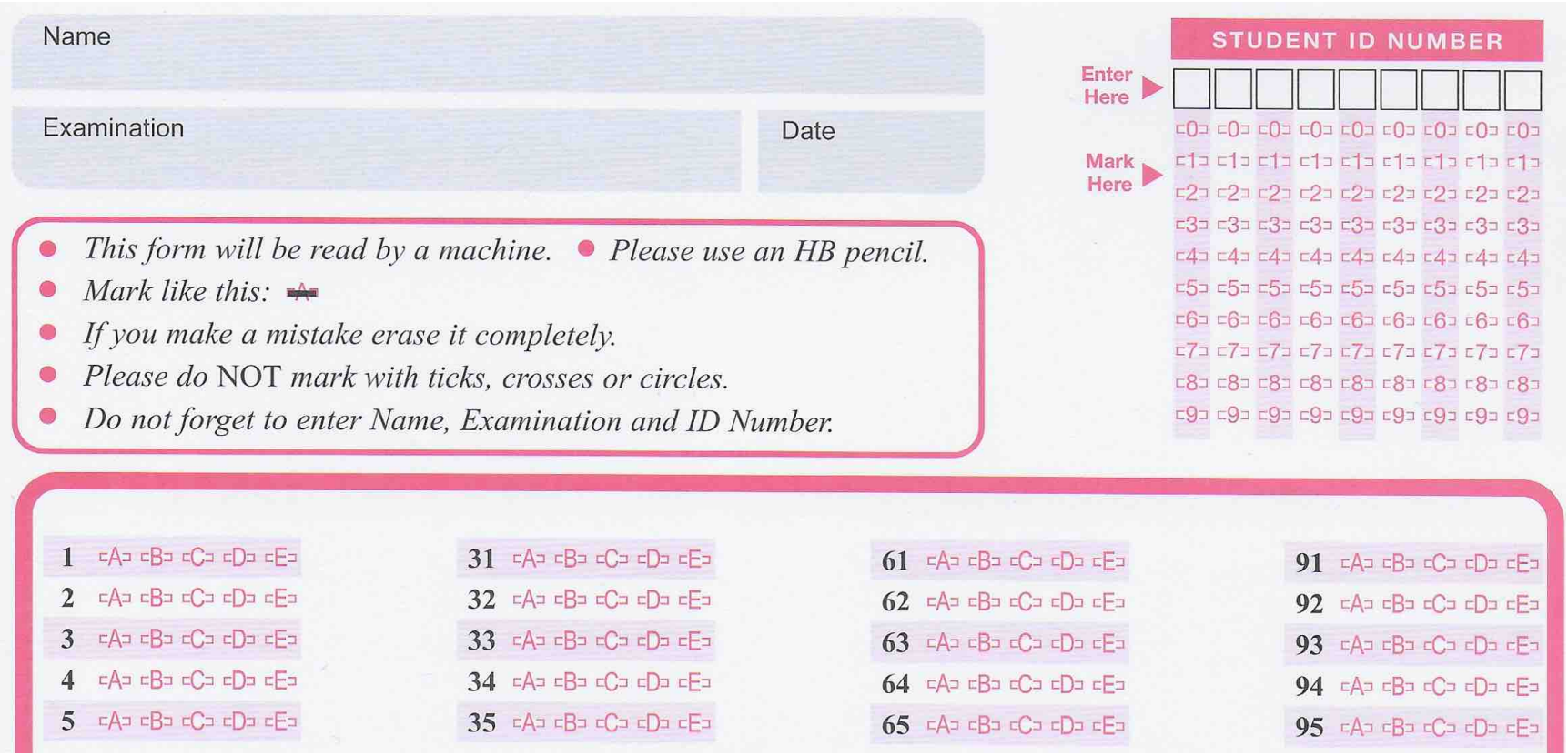
CANDIDATES should enter below in the left hand column the numbers of the questions answered in the order in which they have been attempted.

Question Number	For Examiners use only	

Question Number	For Examiners use only	

TOTAL MARK

The red answer paper for MCQs


Name

Examination

Date

- *This form will be read by a machine.*
- *Please use an HB pencil.*
- *Mark like this:* ~~A~~
- *If you make a mistake erase it completely.*
- *Please do* NOT *mark with ticks, crosses or circles.*
- *Do not forget to enter Name, Examination and ID Number.*

STUDENT ID NUMBER

Enter Here

Mark Here

0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5
6 6 6 6 6 6 6 6 6
7 7 7 7 7 7 7 7 7
8 8 8 8 8 8 8 8 8
9 9 9 9 9 9 9 9 9

1 A B C D E	31 A B C D E	61 A B C D E	91 A B C D E
2 A B C D E	32 A B C D E	62 A B C D E	92 A B C D E
3 A B C D E	33 A B C D E	63 A B C D E	93 A B C D E
4 A B C D E	34 A B C D E	64 A B C D E	94 A B C D E
5 A B C D E	35 A B C D E	65 A B C D E	95 A B C D E

It's ok to mark answers to MCQs by a pen. If you wish, you could bring your pencil for rubbing out. Extra copies of red papers will be available.

Your questions and the quiz

To benefit from the lecture, now you could

- ask or submit your anonymous questions to the COMP229 folder after the lecture;
- write down your summary in 2-3 phrases, e.g. list key concepts you have learned;
- talk to your classmates to revise the lecture.

Question. How can you make sure that your new knowledge acquired during a day is preserved for a long time and won't vanish in the next 8 hours?

Answer to the quiz and summary

Answer. Sleep a lot (hours before midnight are especially valuable), because your brain needs a relaxing phase to strengthen neural connections.

- Most important to pass the module: come to the exam at 10am on Monday 21st Jan 2019.
- During revisions, focus on simpler concepts rather than wasting time on harder topics.
- Revise and also exercise regularly (daily).
- The pass mark is 40%. Your exam marks will likely be forgotten even by you in few months.