Lecture 21. Revision

(the content of this deck is non-examinable)

COMP90051 Statistical Machine Learning

Semester 2, 2018
Lecturers: Ben Rubinstein

This lecture

- Project 2 update
- Exam tips
- Reflections on the subject

Exam Tips

Don’t panic ☺

Exam tips

- Don’t panic!
- Attempt all questions
  - Do your best guess whenever you don’t know the answer
- Finish easy questions first (do q’s in any order)
- Start questions on a new page (not sub-questions)
- If you can’t answer part of the question, skip over this and do the rest of the question
  - you can still get marks for later parts of the question
  - we don’t repeatedly penalise for carrying errors forward
- Answers in point form are fine

What’s non-examinable?

- Green slides
- This deck (well, it’s just a review)
- Note that material covered in the reading is fair-game

- All that said, we won’t put large weight on material given little “air time” → prioritise revision

Changes from previous years

- Last year's exam questions are representative of what you will get at the exam
  - Make sure you understand the solutions!
- Dropped topics in 2017
  - active learning; semi-supervised learning
- New topics in 2017
  - independence semantics in PGMs, HMM details
  - deeper coverage of kernels & basis functions, optimisation, regularisation
- Dropped topics in 2018
  - Manifold learning, spectral clustering, Isomap
  - HMM detail (not dropped, made green)
- New topics in 2018
  - Multi-armed bandits

Exam format

- Four parts A, B, C, D; worth 14, 16, 10, 10 marks
- Total of 50 marks, split into 12 questions
- 180 minutes (3 hours), so 3.6 min / mark
- A = short answer (1-2 sentences, based on #marks)
- B = method questions
- C = numeric / algebraic questions
- D = design & application scenarios

Sample A questions (each 1-2 marks)

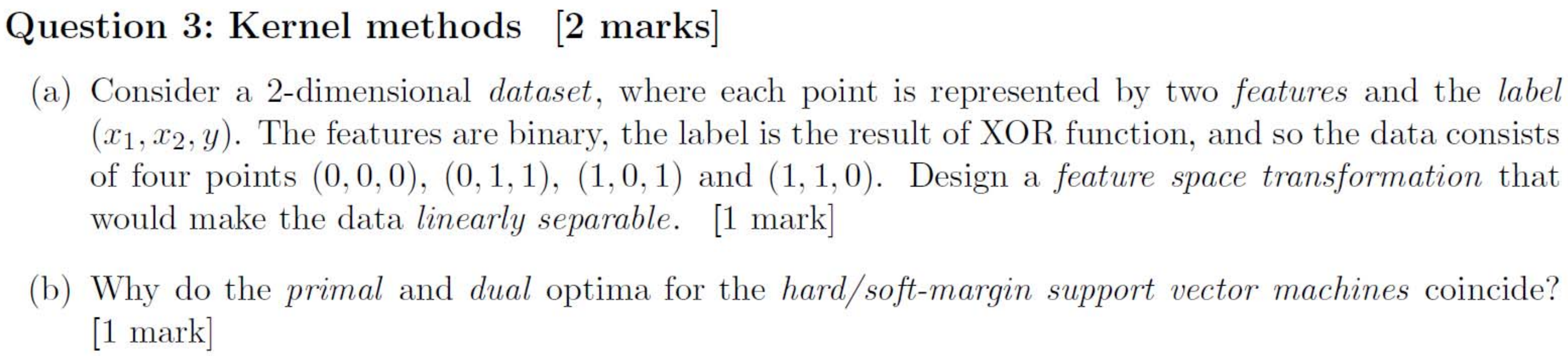
2. In words or a mathematical expression, what is the *marginal likelihood* for a *Bayesian probabilistic model*? [1 mark]

Acceptable: the joint likelihood of the data and prior, after marginalising out the model parameters
Acceptable: $p(\mathbf{x}) = \int p(\mathbf{x}|\theta)p(\theta)d\theta$ where $\mathbf{x}$ is the data, θ the model parameter(s), and $p(\mathbf{x}|\theta)$ the likelihood and $p(\theta)$ the prior
Acceptable: the expected likelihood of the data, under the prior

4. In words, what does $\Pr(A, B \mid C) = \Pr(A \mid C)\Pr(B \mid C)$ say about the *dependence* of A, B, C? [1 mark]

A and B are conditionally independent given C.

Sample B question (each 2-4 marks)



Question 3: Kernel methods [2 marks]

(a) Consider a 2-dimensional *dataset*, where each point is represented by two *features* and the *label* (x_1, x_2, y). The features are binary, the label is the result of XOR function, and so the data consists of four points $(0, 0, 0)$, $(0, 1, 1)$, $(1, 0, 1)$ and $(1, 1, 0)$. Design a *feature space transformation* that would make the data *linearly separable*. [1 mark]

(b) Why do the *primal* and *dual* optima for the *hard/soft-margin support vector machines* coincide? [1 mark]

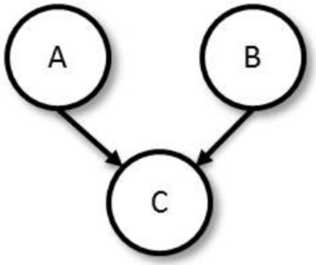
Acceptable: new feature space (x_3), where $x_3 = (x_1 - x_2)^2$

Acceptable: As strong duality holds due to convexity of the objectives.

Sample C question (each 2-4 marks)

Question 5: Statistical Inference [3 marks]

Consider the following directed PGM



where each random variable is Boolean-valued (True or False).

1. Write the format (with empty values) of the conditional probability tables for this graph. [1 mark]
2. Suppose we observe n sets of values of A, B, C (complete observations). The maximum-likelihood principle is a popular approach to training a model such as above. What does it say to do? [1 mark]
3. Suppose we observe 5 training examples: for (A, B, C) — $(F, F, F); (F, F, T); (F, T, F); (T, F, T); (T, T; T)$. Determine maximum-likelihood estimates for your tables. [1 mark]

Sample C question (cont)

1. CPTs [1 mark]

```
------------   ------------   ------------------
Pr(A=True)     Pr(B=True)     A B Pr(C=True|A,B)
------------   ------------   ------------------
?              ?              T T ?
------------   ------------   T F ?
                              F T ?
                              F F ?
                              ------------------
```

2. MLE [1 mark]

Acceptable: It says to choose values in the tables that maximise the likelihood of the data.
Acceptable: $\arg\max_{tables} \prod_{i=1}^{n} \Pr(A = a_i) \Pr(B = b_i) \Pr(C = c_i \mid A = a_i, B = b_i)$

3. Show MLE [1 mark]

The MLE decouples when we have fully-observed data, and for discrete data as in this case — where the variables are all Boolean — we just count.
The $Pr(A = True)$ is 2/5 since we observe A as true out of five observations. Similarly for B we have the probability of True being 2/5. Finally for each configuration TT, TF, FT, FF of AB we can count the times we see C as True as a fraction of total times we observe the configuration. So we get for these probability of $C = True$ as 1.0, 1.0, 0.0, 0.5 respectively.

Reflections on the Subject

What is Machine Learning?

- Machine learning
  - *“a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (such as planning how to collect more data!)”* (Murphy)
- Data mining
- Pattern recognition
- Statistics
- Data science
- Artificial intelligence

Common themes

- Different schools of thoughts (frequentist, Bayesian, decision theoretic) differ in how parameters are modelled or fit/learned
- Most approaches are probabilistic or loss-based
- Many roads lead to the same algorithms
- Regularisation, model selection, over/underfitting
- The importance of scratching the broad surface

Thank you and good luck!