Lecture 25. Revision

(the content of this deck is non-examinable)

COMP90051 Statistical Machine Learning

Semester 2, 2017
Lecturers: Trevor Cohn

THE UNIVERSITY OF MELBOURNE

This lecture

- Project wrap-up
- Exam tips
- Reflections on the subject
- Q&A session

Project 2

Well done everyone!

SVHN: House Numbers from photos

Taken from Google Street view images

- Manual bounding boxes by AMT workers
- Becoming a new standard benchmark problem, following MNIST
- 200k images, about 600k digits
- Varying resolution

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. (2011). *Reading digits in natural images with unsupervised feature learning*. Deep Learning and Unsupervised Feature Learning Workshop, NIPS.

Processing pipeline

1. Extract images from bounding boxes for each digit
2. Normalise colours
3. Flatten to greyscale
4. Filter out instances with low contrast
5. Resize to 64x64 dimensions

Official dataset: SOTA ~90%, human 98%

Kaggle rankings

- Often large change in ranking vs public leaderboard
- Scored based on ranking, where ties assigned equal rank

#	△pub	Team Name	Score
1	▲68	JimmyChiang	1.00000
2	▲19	BoomW	0.99599
3	▲53	Rango	0.99199
4	▲53	nekketsuing	0.99199
5	▼2	Sheng	0.99199
6	▲10	HHHHH	0.99199
7	▲60	wfymai	0.99199
8	▲26	EE V587	0.99199
9	▼4	Exponential Family	0.99199
10	▲42	lligtt	0.99199
11	▲77	WillFang	0.98799
12	▼11	xieyue	0.98799
13	▼6	NoNeedANameLoseA...	0.98799
14	▲12	bryan	0.98799
15	▼1	MinG	0.98799
16	▲82	ying Li	0.98799
17	▲45	Lemonade	0.98799
18	▲7	mathias	0.98799
19	▼7	Emmmmmm	0.98799
20	▲45	reubenv	0.98799

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 multiply 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)
- Something that was in workshops but not in lectures
- Note that material covered in the reading is fair-game

Changes from last year

- 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

Exam format

- Four parts A, B, C, D; worth 13, 17, 10, 10 marks
- Total of 50 marks, split into 11 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 3-6 marks)

Question 3: Kernel methods [2 marks]

1. 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]

2. Intuitively what does the *Representer Theorem* say? [1 mark]

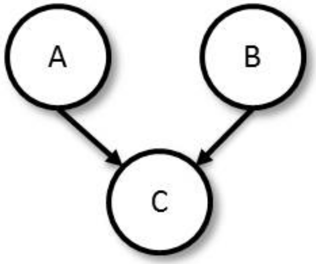
Acceptable: new feature space (x_3), where $x_3 = (x_1 - x_2)^2$

Acceptable: a large class of linear models can be formulated such that both training and making predictions require data only in a form of a dot product
Acceptable: The solution to the SVM (the weight vector) lies in the span of the data.
Acceptable: $\mathbf{w}^\star = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$ or something similar.

Sample C question (each 2-3 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.

A Deeper Insight

A selection of additional topics
with the aim to provide a deeper
insight into main lectures content

Networks in real life: the Internet

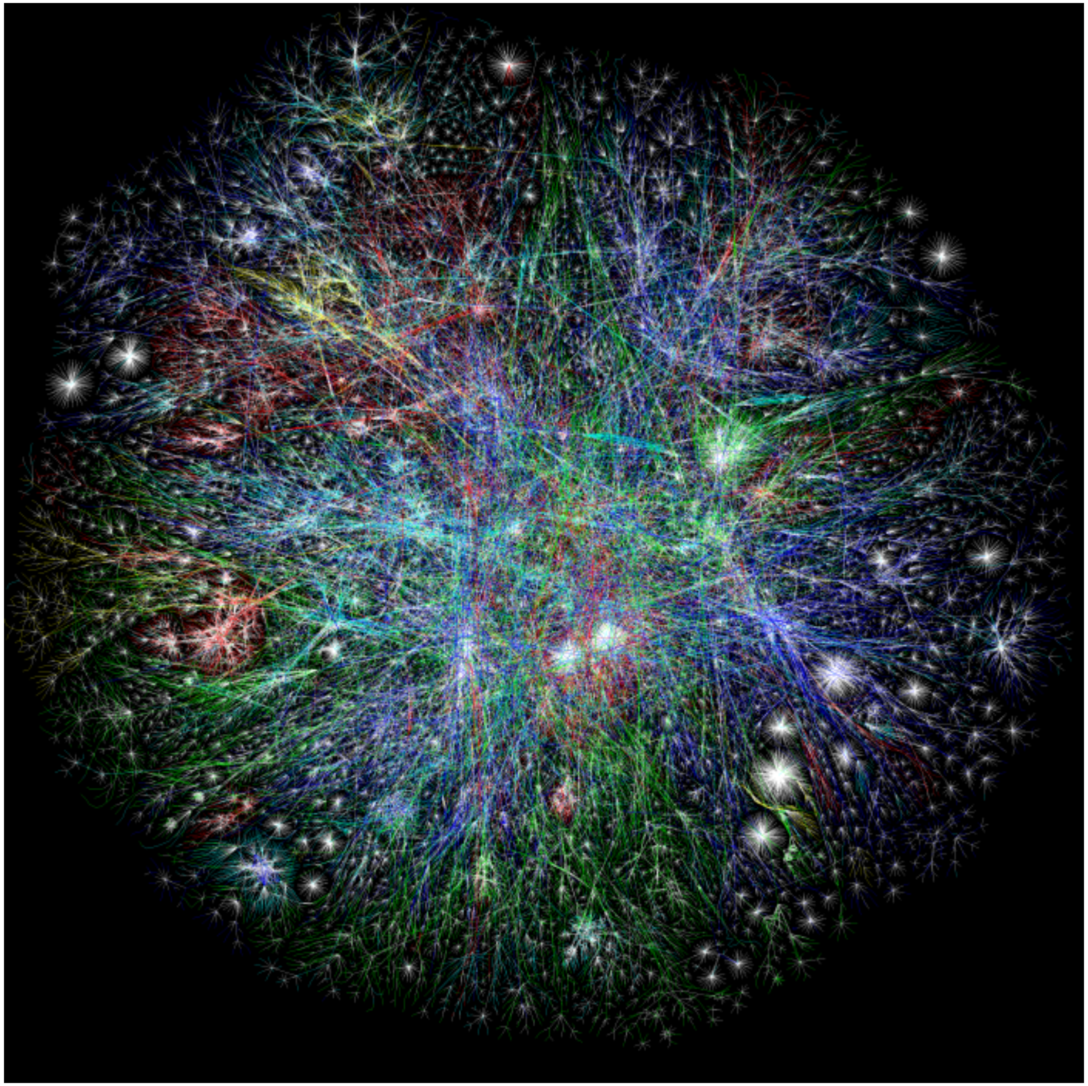
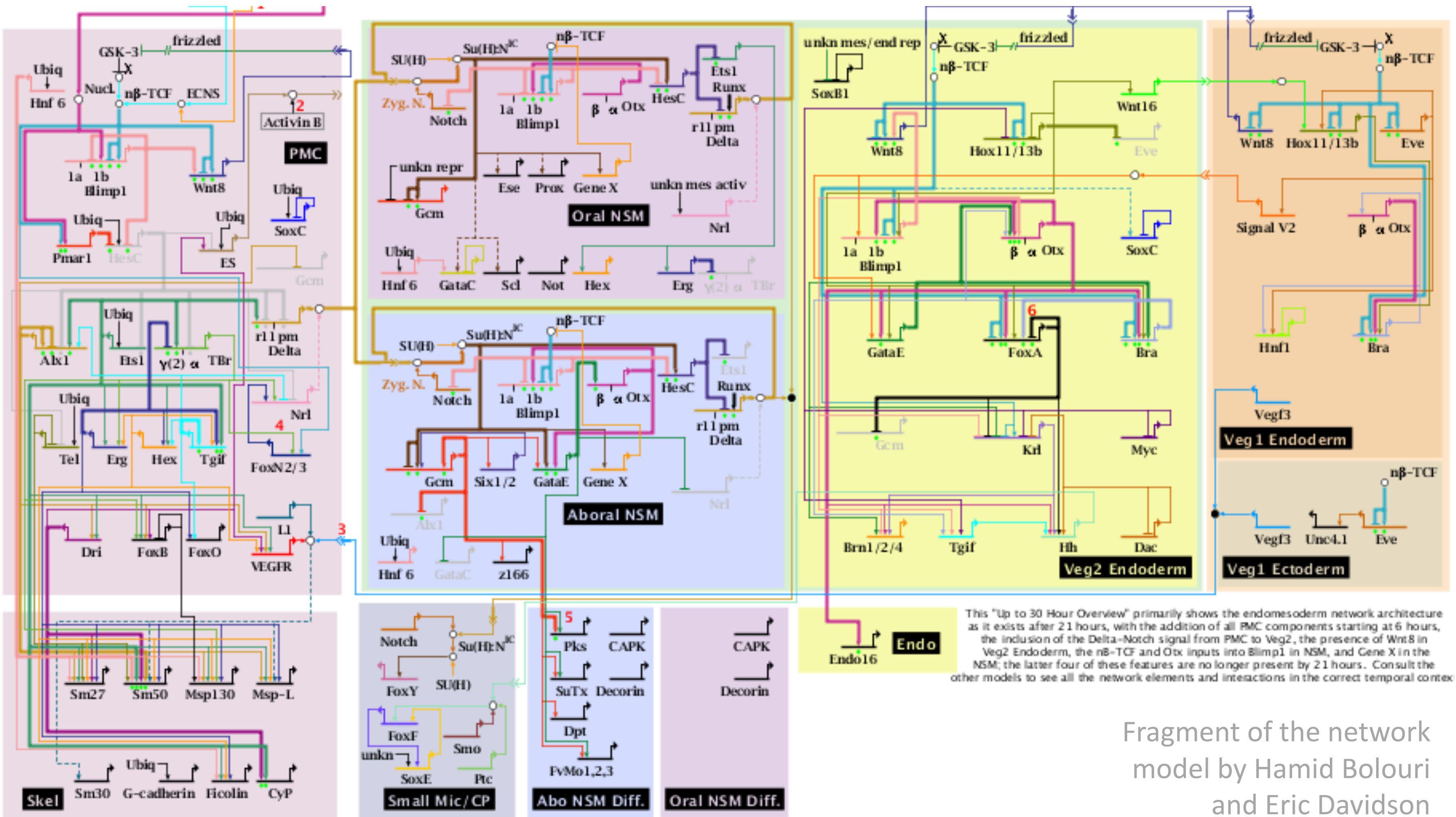


Image: OPTE Project Map (CC2)

Networks in real life: gene regulatory network



Fragment of the network model by Hamid Bolouri and Eric Davidson

Networks in real life: transport map

Network analysis (1/4)

- Analysis of large scale real world networks has recently attracted considerable attention from research and engineering communities
- Networks/graphs is a list of pairwise relations (edges) between a set of objects (vertices)
- Example problems / types of analysis
  * Link prediction
  * Identifying frequent subgraphs
  * Identifying influential vertices
  * Community finding

Network analysis (2/4)

- *Community is a group of vertices* that interact more frequently within its own group than to those outside the group
  - Families
  - Friend circles
  - Websites (communities of webpages)
  - Groups of proteins that maintain a specific function in a cell
- This is essentially a definition of a *cluster* in unsupervised learning

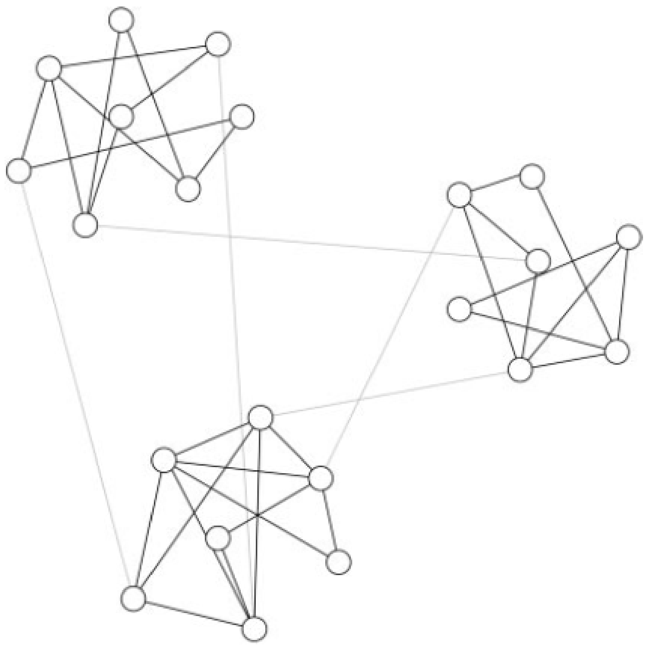


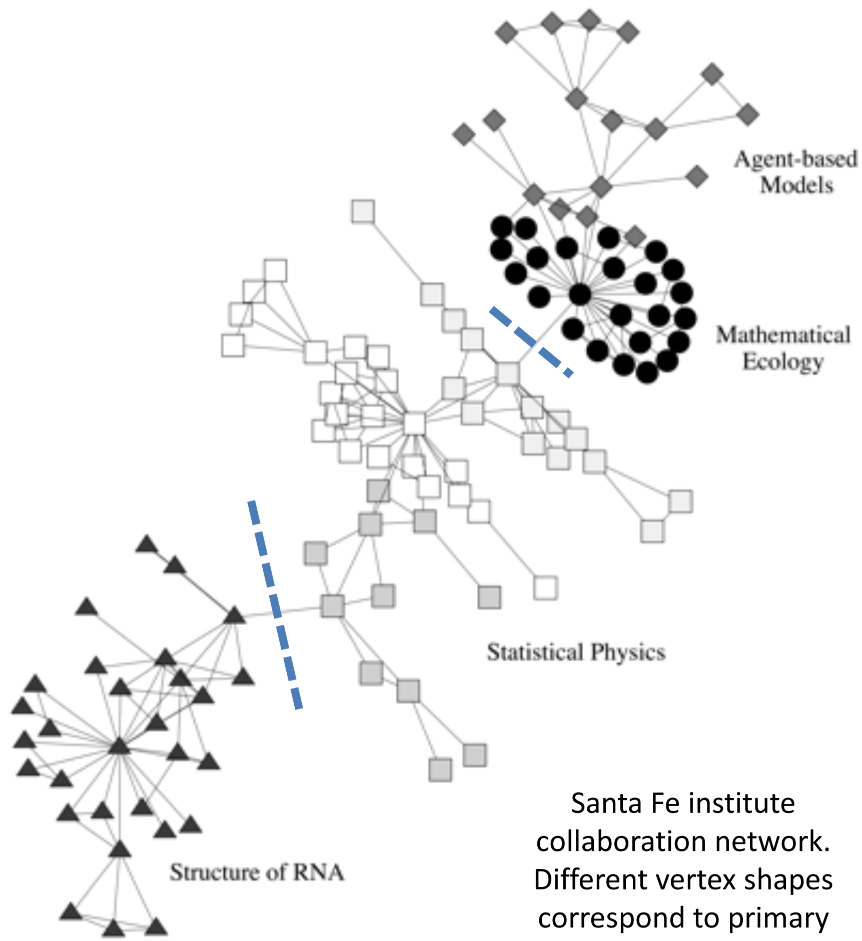
Image: Girvan and Newman, Community structure in social and biological networks, PNAS, 2002

Network analysis (3/4)

- Why community detection?
  - Understanding the system behind the network (e.g., structure of society)
  - Identifying roles of vertices (e.g., hubs, mediators)
  - Summary graphs (vertices – communities, edges – connections between communities)
  - Facilitate distributed computing (e.g., place data from the same community to the same server or core)
- There are many community detection algorithms, let’s have a look at only one of the ideas

Network analysis (4/4)

- Communities are connected by a few connections, which tends to form *bridges*
- Cut the bridges to obtain communities
- One of the algorithms is called normalised cuts which is equivalent to spectral clustering



Santa Fe institute collaboration network. Different vertex shapes correspond to primary divisions of the institute

Image: Girvan and Newman, Community structure in social and biological networks, PNAS, 2002

Reflections on the Subject

Supervised learning

- Essentially a task of function approximation
- A function can be defined
  * Theoretically, by listing the mapping
  * Algorithmically
  * Analytically
- Every equation is an algorithm, but not every algorithm is an equation

Supervised learning

- Simple and more interpretable methods (e.g., linear regression) vs more complicated “black box” models (e.g., random forest)
- Apparent dichotomy: prediction quality vs interpretability
- However, some complex models are interpretable
  - Convolutional Neural Networks
  - In any “black box” model, one can study effects of removing features to get insights what is a useful feature

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

I’ll first stay here, then move
to the office hour room

Thank you and good luck!