Inf2b - Learning

Lecture 16: Review

Hiroshi Shimodaira
(Credit: Iain Murray and Steve Renals)

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

http://www.inf.ed.ac.uk/teaching/courses/inf2b/
https://piazza.com/ed.ac.uk/spring2020/infr08028
Office hours: Wednesdays at 14:00-15:00 in IF-3.04

Jan-Mar 2020

Today's Schedule

1. Topic revision
2. Maths formulae to remember
3. Methods/derivations to understand
4. Exam technique

Topics dealt within the course

- Distance and similarity measures (Pearson correlation coef.)
- Clustering (K-means clustering)
- Dimensionality reduction (covariance matrix, PCA)
- Classification
  - K-NN classification
  - Naive Bayes
  - Gaussian classifiers (MLE, discriminant functions)
  - Neural networks (Perceptron error correction algorithm, sum-of-squares error cost function, gradient descent, EBP)
- Statistical pattern recognition theories
  - Bayes theorem, and Bayes decision rule
  - Probability distributions and parameter estimation
    - Bernoulli distribution / Multinomial distribution
    - Gaussian distribution
  - Discriminant functions
  - Decision boundaries/regions (minimum error rate classification)
  - Evaluation measures and methods
- Optimisation problems

Maths formulae to remember

- Euclidean distance:

$$r_2(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}|| = \sqrt{\sum_{i=1}^{D}(x_i - y_i)^2}$$

cf. $\mathrm{sim}(\mathbf{x}, \mathbf{y}) = \frac{1}{1+r_2(\mathbf{x},\mathbf{y})}$ as a similarity measure

- Pearson correlation coefficient:

$$\rho(x, y) = \frac{1}{N-1}\sum_{n=1}^{N}\frac{(x_n - \mu_x)}{\sigma_x}\frac{(y_n - \mu_y)}{\sigma_y}$$

- Bayes Theorem

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$$P(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)P(C_k)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|C_k)P(C_k)}{\sum_{k=1}^{K} p(\mathbf{x}|C_k)P(C_k)}$$

Maths formulae to remember *(cont.)*

- Bayes decision rule (cf. MAP decision rule)

$$k^* = \arg\max_k P(C_k \,|\, \mathbf{x}) = \arg\max_k P(\mathbf{x}|C_k)P(C_k)$$

- Naive Bayes for document classification

(vocabulary: $V = \{w_1, \ldots, w_{|V|}\}$, test document: $D = (o_1, \ldots, o_L)$)

  - Likelihood by Bernoulli document model

$$P(\mathbf{b}|C_k) = \prod_{t=1}^{|V|}[b_t P(w_t \,|\, C_k) + (1-b_t)(1-P(w_t \,|\, C_k))]$$

  - Likelihood by Multinomial document model

$$p(\mathbf{x}|C_k) \propto \prod_{t=1}^{|V|} P(w_t|C_k)^{x_t} = \prod_{i=1}^{L} P(o_i|C_k)$$

Maths formulae to remember *(cont.)*

- Univariate Gaussian pdf:

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

- Multivariate Gaussian pdf:

$$p(\mathbf{x}\,|\,\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Parameter estimation from samples:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n, \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n - \hat{\boldsymbol{\mu}})(\mathbf{x}_n - \hat{\boldsymbol{\mu}})^T$$

NB: N in case of MLE

- Correlation coefficient:

$$\rho(x_i, x_j) = \rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}, \qquad \Sigma = (\sigma_{ij})$$

Maths formulae to remember *(cont.)*

- Logistic sigmoid function:

$$y = g(a) = \frac{1}{1 + \exp(-a)}$$

$$g'(a) = g(a)(1-g(a))$$

- Softmax activation function (for multiple output nodes):

$$y_k = \frac{\exp(a_k)}{\sum_{\ell=1}^{K}\exp(a_\ell)}$$

- and basic maths rules (e.g. differentiation)

Methods/derivations to understand (non exhaustive)

- Clustering and classification
- Discriminant functions of Gaussian Bayes classifiers
- Learning as an optimisation problem
  - Maximum likelihood estimation
  - Gradient descent and back propagation algorithm (neural networks) for minimising the sum-of-squares error

NB: Learning is a difficult problem by nature — generalisation from a limited amount of training samples.
→ need to assume some structures (constraints):
  - Probability distributions
  - Naive Bayes
  - Diagonal covariance matrix rather than a full covariance for each class, shared covariance matrix among classes, regularisation.
  - Dimensionality reduction and feature selection *(NE)*

Machine learning as optimisation problems

- Euclidean-distance based classification

$$k^* = \arg\min_k \|\mathbf{x} - \mathbf{r}_k\|$$

- K-means clustering

$$\min_{\{z_{kn}\}}\sum_{k=1}^{K}\sum_{n=1}^{N} z_{kn}\,\|\mathbf{x}_n - \mathbf{m}_k\|^2$$

- Dimensionality reduction to 2D with PCA

$$\max_{\mathbf{u},\mathbf{v}} \mathrm{Var}\,(y) + \mathrm{Var}\,(z)$$

subject to $\|\mathbf{u}\| = 1, \|\mathbf{v}\| = 1, \mathbf{u} \perp \mathbf{v}$

- Bayes decision rule

$$k^* = \arg\max_k P(C_k\,|\,\mathbf{x}) = \arg\max_k P(\mathbf{x}|C_k)P(C_k)$$

- Maximum likelihood parameter estimation

$$\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{D})$$

- Least squares error training of neural networks

$$\min_{\mathbf{w}} \frac{1}{2}\sum_{n=1}^{N}\|\mathbf{y}_n - \mathbf{t}_n\|^2$$

Exam revision

Look at lecture notes, slides, tutorials, coursework, and past papers.

Early exam papers: many (useful) multiple choice Qs
- No longer the exam format
- Syllabus has changed slightly

Recent exam papers since 2008/09
- Answer two questions from section A (ADS) and two questions from section B (Learning).
- Closed-book exam.
- Calculators may be used (approved ones only).
- Solutions are available only for 2008/09, 2009/10, 2013/14 (no plans of releasing those of missing years)
- NB: errors in some solutions, e.g. 5 (c) of 2008/09: square root is not taken in computing standard deviations.

Well prepared for the exam of 120 minutes

60 minutes/section, 30 minutes/question

Exam revision *(cont.)*

Don't overfit!

Anything that appears in the notes, slides, tutorial sheets, or coursework is examinable, unless marked non-examinable, extra topics, or (†)

Don't trust unofficial solutions

Inf2b Revision Meeting
- Date: TBC (in late April)
- Send me questions/requests that you want me to discuss at the meeting.

Time in the exam

- Half an hour per question (minus time to pick questions)
- Don't panic!
- Go for easy marks first
- Don't spend a long time on any small part
- Don't scrawl - you might lose marks if the marker cannot read/understand
- Know the standard stuff:

  there's not time to work everything out from scratch

Calculators may be used in the examination: The School of Informatics does not provide calculators for use in exams. If the use of a calculator is permitted in an exam, it's your responsibility to bring an approved calculator to the exam.

End-of-course feedback:

Thanks!