Inf2b Learning and Data

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/class#spring2016/infr08009learning

Jan-Mar 2016

Today's Schedule

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

Topics dealt within the course

- Distance and similarity measures (collaborative filtering)
- Clustering (K-means clustering)
- Classification
  - K-NN classification
  - Naive Bayes
  - Gaussian classifiers (maximum-likelihood estimation, discriminant functions)
  - Neural networks (Perceptron error correction algorithm, sum-of-squares error cost function, gradient descent, error back propagation)
- 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
  - Evaluation measures and methods

Maths formulae to memorise

- Euclidean distance:
$$r_2(\boldsymbol{x},\boldsymbol{y}) = ||\boldsymbol{x}-\boldsymbol{y}|| = \sqrt{\sum_{i=1}^{D}(x_i-y_i)^2}$$
cf. $\mathrm{sim}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{1+r_2(\boldsymbol{x},\boldsymbol{y})}$ 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
$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$
$$P(C_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_k)P(C_k)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|C_k)P(C_k)}{\sum_{k=1}^{K}p(\boldsymbol{x}|C_k)P(C_k)}$$

Maths formulae to memorise (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
$$P(\boldsymbol{b}|C_k) = \prod_{t=1}^{|V|}[b_tP(w_t\,|\,C_k)+(1-b_t)(1-P(w_t\,|\,C_k))]$$
  - 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 memorise (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:
$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\,, \qquad \hat{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n-\hat{\boldsymbol{\mu}})(\mathbf{x}_n-\hat{\boldsymbol{\mu}})^T$$
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 memorise (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):
  - Naive Bayes
  - Diagonal covariance matrix rather than a full covariance for each class, shared covariance matrix among classes, regularisation.

Machine learning as optimisation problems

- Euclidean-distance based classification
$$k^* = \arg\min_k \|\boldsymbol{x}-\mathbf{r}_{C_k}\|$$
- K-means clustering
$$\min_{\{z_{kn}\}}\frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N}z_{kn}\,\|\mathbf{x}_n-\mathbf{m}_k\|^2$$
- 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
$$\min_{\boldsymbol{w}}\frac{1}{2}\sum_{n=1}^{N}\|\mathbf{y}_n-\mathbf{t}_n\|^2$$

Exam revision

Look at lecture notes, slides, tutorials, 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

- Solutions are available only for 2008/09, 2009/10, 2013/14 (no plans to release those of missing years)
- NB: error in the solution for 5 (c) of 2008/09: square root is not taken in computing standard deviations.

Don't overfit!

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

Don't trust unofficial solutions

Exam revision (cont.)

- There will be an Inf2b Revision Meeting in April before the exam
  - Date: TBC
  - 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
- 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.