Learning from Data: Data and Models

Amos Storkey, School of Informatics

September 26, 2005

http://www.anc.ed.ac.uk/~amos/lfd/

Tutorials

- Please sign up: http://www.anc.ed.ac.uk/∼amos/lfd/tutorform.html
- 1 tutorial a week.
- Tutorials start Thurs week 3.
- Also you MUST sign up with the Teaching Office
- http://www.inf.ed.ac.uk/admin/ITO/registration/sem1-registration.html

Data? What You See is What You Get

- Data types:
  - Real valued, positive, vector (geometric), bounded, thresholded.
  - Categorical data, hierarchical classes, multiple membership, etc.
  - Ordinal data, binary data, partially ordered sets.
  - Mixed, scale related, scale unrelated, hierarchical importance.
  - Missing, with known error, with error bars (known measurement error).
  - Internal dependencies, conditional categories.
  - Raw, preprocessed, normalised, transformed etc.
  - Biased, corrupted, just plain wrong, in unusable formats.
  - Possessed, promised, planned, non-existent.

Attributes and Values

- Datasets can be thought of as attribute value pairs.
- For example 'state of weather' is an attribute and 'raining' is a value.
- 'Height' is an attribute, and '4ft 6in' is a value.
- For analysis purposes it is handy to re-represent values as numerical quantities.

Example Attribute Possibilities

- 1,2,3,4,5,6,7,8,9,10
- Red, Blue, Green, Yellow, Pink
- 1.7556, 3.449432, 2.34944, etc
- Strongly Disagree, Disagree, Neither Disagree or Agree, Agree, Strongly Agree
- ...,-3,-2,-1,0,1,2,3,...

Categorical Data

- Each observation belongs to one of a number of categories. Orderless.
- Example: type of fruit (orange, apple, pear, grape).
- one-of-m encoding. Represent each category by a particular component of an attribute vector. Eg orange= (1000), apple= (0100), pear= (0010) and grape= (0001). In other words each attribute is a indicator function for the particular fruit type.
- Note only one component can be 'on' at any one time. This means that the attributes cannot be independent.

Ordinal Data

- Each observation belongs to one of a number of categories. Ordered.
- Example: university grade: (3,2ii,2i,1).
- Numeric encoding. Represent each category by a number. Keep to correct order.
- Note that models which use the assigned numeric values rather than just the order of the values are not really respecting the data. Keep in mind the arbitrary nature of the numeric allocation.

Numeric Data

- Integers or real numbers.
- Numeric values mean something. It is meaningful to add, multiply etc.
- Integer valued variables can not always be treated in the same way as real valued variables.

Thinking Generatively

- How might the data have been generated:
  - Hidden variables. Independence assumptions.
- What key characteristics do we believe the data might have?
  - Smoothness.
  - Relational structure.
  - Linearity.
  - A particular distribution.
  - Noise levels?
  - Low dimensionality (degrees of freedom).
  - Multiple sources.

Examples

- Brightness of astronomical objects.
  - Stars and Galaxies quite different.
  - Propose a two source model - different models for brightness for each source type.
- position of 40 points on a robot arm.
  - Robot arm only has a few degrees of freedom - expect a lower dimensional representation.
- Loan defaulting.
  - Expect people in similar circumstances to be at similar risk of defaulting.

Relevant Variables

- Variable selection, feature selection.
- What are the things which the variables we are interested in is dependent on?
  - Variables that may affect it.
  - Variables that it may affect.
  - Variables that indirectly affect it.
  - Statistics (features) calculated from variables which are useful information.

Learning?

- So what is this course about?
- Analogy with human learning.
  - Learning: spending time obtaining general knowledge about the world around us.
  - Inference: Using our knowledge to work out what is going on in a specific scenario, or regarding a specific instance.
- (In fact learning is just the same as inference, but about more general information.)
- Learn models from given data. Use models to infer things about new data.

Dasher

- Dasher learns a model for the patterns which occur in texts.
- Can train with your own past writing.
- Now given what you have ‘typed’ (specific instance) need to work out what you are about to ‘type’.
- Learn models from given data. Use models to infer things about new data.

Supervised or Unsupervised?

- Supervised learning: have training data to help learn the model.
- Usually data takes form of input-output values. (page 4-5 of notes).
- Unsupervised learning: just have a set of examples - want to learn the characteristics of that set.
- No outputs or inputs. (page 4 of notes).
- How is the data going to be used? For test purposes are you trying to predict the value of one or more fields given the others (supervised)?
- Or are you wanting to make more general statements about the characteristics of the data as a whole (unsupervised)?

Generative or Discriminative?

- Generative: try to model the distribution of the whole of the data.
- Similar to unsupervised.
- Discriminative: no requirement to model the distribution of much of the data. Only interested in how one particular part of the data (eg one field), depends on the others.
- Similar to unsupervised. Difference is the focus here is on the form of *model*.
- Any generative approach can be used discriminatively, but not vice versa. But good generative modelling much harder.

Summary

- Data is messy. Can take many forms. Can be many practical problems.
- Real, categorical, ordinal.
- Thinking generatively.
- Learning and Inference.
- Supervised and Unsupervised. Generative and Discriminative.