Thinking about Data Quiz

1. ML covered in IAML
   - classification
   - regression
   - clustering

2. ML problem
   - large DB of prev. opinion polls and election results, learn a function that can predict a forthcoming election result given current opinion polls

3. Numeric attribute - 1.1, 2.3, 4.6
   Ordinal - Low, Medium, High
   Categorical - Animal, Vegetable, Fish

4. Categorical : $a = b$

5. Ordinal : $a < b$, $a = b$ (equality and ordering)

6. Numerical: $a + b$, $a < b$, $a = b$

7. Only numeric atts can be normalised

8. Purpose of normalisation
   - make values of attributes (different) be roughly comparable

9. Normalisation methods
   - converting to range $[0, 1]$
   - converting to zero mean and unit variance

10. Outliers can be found by visualising and observing d.ps w/ unusual values
    Skewed distributions can be dealt by taking the log. of values
    A non-monotonic relationship between numerical/ordinal values and the class can be dealt by converting values to a variety of overlapping numerical ranges.

11. Normalising (e.g. if we scale, centre and deslant the image) allows us to use post-normalisation pixel values as features for classification.
    ⇒ depends on the types of images and the classification task.

12. For object recognition in images, we want attributes that are: (rotation, ✓)
    - invariant to irrelevant differences in the training images (scale, lighting, etc.)
    - have similar values for images in the same class and diff. values for images in different classes
      - spread of values among classes → easy to distinguish

13. text representation
- use the possible words (vocabulary) as the attrs and the existence of that word in the document as the value. → word count

14. sound signals (speech, music, etc.) representation
- frequencies present in signal

15. Accuracy is a good success metric for comparing classifiers if
- the classes are balanced.

16. Generative - models classes
- almost always probabilistic
- can use unlabeled data

Discriminative - models decision boundary
- may/may not be probabilistic
- can't use unlabeled data