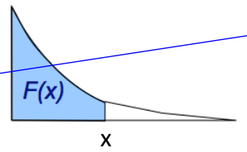
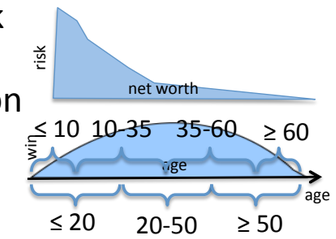
Numeric attributes: issues

- Skewed distributions
  - systematic extreme values
  - affects regression, kNN, NB; but not DTs
  - simple fix: log(x) or atan(x), then normalize
  - cumulative distribution function: x' = F(x) = P(X≤x)



- Non-monotonic effect of attributes
  - affects regression, NB, DTs(gain); less important for kNN
  - monotonic: net worth and lending risk
    - higher net worth → lower lending risk
  - non-monotonic: age → win a marathon
    - sweet spot: not too young, not too old
  - simple fix: quantization
    - can be unsupervised, overlapping



Overview

- Attribute-value pairs
- Examples of real data
  - credit scoring
  - handwritten digits
  - object recognition
  - text classification
- Issues to consider

Comments on page 7

How does it affect regression? My understanding was that in regression the distribution of the residuals is assumed to be symmetric, but no assumptions are made about the distributions of the attributes - only that the relationship between the attributes and the dependent variable is linear (assuming we're doing linear regression). So does an attribute having a skewed distribution *necesarily* affect regression, or is it just that *if* the relationship between the attributes and the dependent variable is non linear, making the distribution of an attribute less skewed might help?

In the same way as outliers. Please see the following examples in the lectures:

http://nb.mit.edu/f/17452?p=12 (Mean Squared Error)

http://nb.mit.edu/f/17812?p=21 (Least Squares Classification)

I get that the mean squared error and least squares estimate will be affected if the distribution of the error terms is skewed or contains outliers. But my point is that as far as I can see, an individual *predictor* having a skewed distribution doesn't necesarily mean that the distribution of the *errors* will be skewed - for example if we have one predictor attribute and one class attribute, and BOTH have heavily skewed univariate distributions, then they could well have a strongly linear relationship with all the distances between true and predicted values being small - right?

First, it's not the class value you are predicting in regression., but that's a minor point.

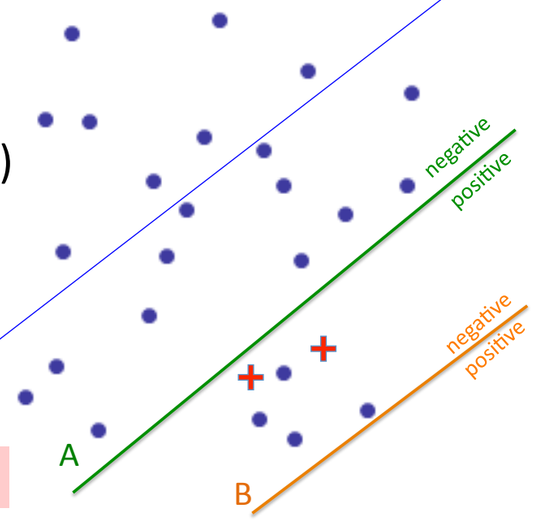
When you're doing regression the assumption is that most of the variance in the dependent variable (the target) can be explained by the independent variable (the attribute). The remaining variance (the noise) is usually fairly small, or else it starts to pull on the model.

With skewed distributions, you are very likely to have large residual variance around the instances with large values, which will affect the quality of the fit a lot more than useful variance around the points wi

By points wi here do you mean the rest of the points when we leave out those with large values? What is more do we want to "fix" skewed distributions of both the dependent and the independent variables? Thank you.

Accuracy and un-balanced classes

- You're predicting Nobel prize (+) vs. not (•)
- Would you prefer classifier A or B?
- Is accuracy (% correct) higher for A or B?
- Accuracy / error rate poor metric here
- Want:
  - cost (Miss) > cost (FA)



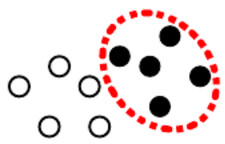
Comments on page 15

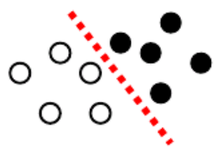
Could you please explain this? Thank you.

The basic accuracy measure assigns equal importance to Misses (False Negatives) and False Alarms (False Positives). If Misses were more important, classifier A would be preferred to classifier B (which is not the case if we use accuracy to evaluate them). We will discuss this in a lot more detail in a few lectures.

Generative vs. Discriminative

- Generative:
  - probabilistic "model" of each class
  - decision boundary:
    - where one model becomes more likely
  - natural use of unlabeled data
- Discriminative:
  - focus on the decision boundary
  - more powerful with lots of examples
  - not designed to use unlabeled data
  - only supervised tasks

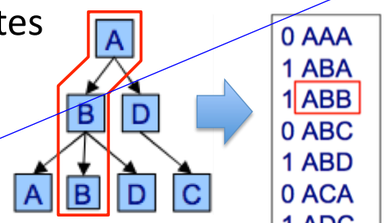




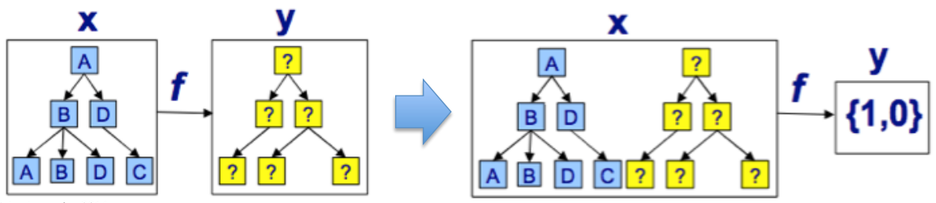
Comments on page 15

Dealing with structure

- Structured input: embed in attributes
  - e.g. tree w. free branching, labels
    - meaning of "A" depends on level
    - one possible representation:
      - attributes = root-to-leaf paths

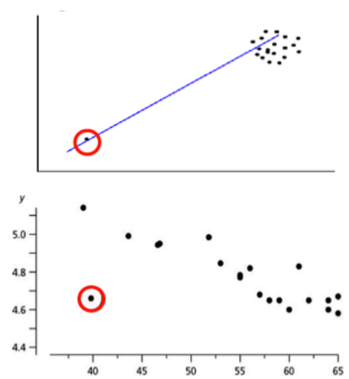


- Structured output: embed in input
  - predict 1/0: output does / doesn't go with input
  - search over possible outputs becomes main focus



Outliers in the data

- Isolated instances of a class that are unlike any other instance of that class
  - affect all learning methods to various degrees
- Extreme attribute values:
  - detect: confidence interval
  - remove or threshold
- Dissimilar to other instances
  - remove or try to fix (mis-labeled?)
- Always try to visualize the data
  - helps detect many irregularities



Comments on page 16

Comments on page 16

Could you give an example of where this would be used?

Thanks!

Machine translation, where you are trying to decide if a hypothetical English sentence Y is a valid translation for a given French sentence X.

Speech recognition: is an English transcription Y a good representation of the audio signal Y.