Regression-based analysis

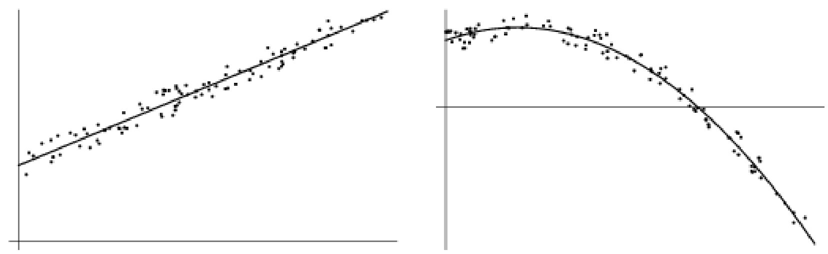
- Prediction versus classification
  - Both construct a model, and use it to predict a value
  - Classification predicts a categorical class label
  - Prediction models continuous-valued functions
- Major method is regression
  - Linear and multiple regression
  - Nonlinear regression (arbitrary curves)

Metric approximation

- Metric: a "distance" function
  - $d(x, x) = 0$
  - $d(x, y) > 0$ iff $x \neq y$ (nontriviality)
  - $d(x, y) = d(y, x)$ (symmetry)
  - $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)
- Metric approximation
  - Find object X in target class C closest to item I
  - X such that for each Y in C, $d(I, X) \leq d(I, Y)$
  - C = curves: regression, curve fitting
  - C = clusters: clustering methods
  - C = patterns: pattern matching

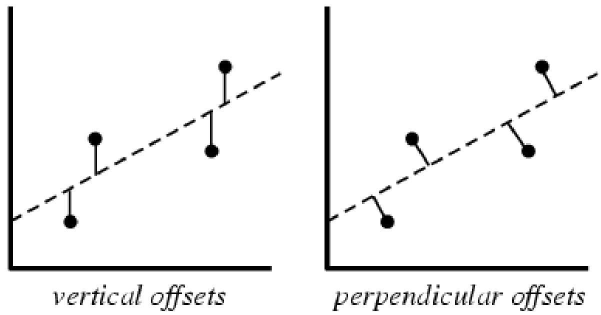
Regression

- Fitting a line or curve to data



Regression approximation

- Ideal metric: straight-line distance (offsets) of points to curve
- Practical approximate metric: distance of vertical offsets



Linear regression

- Linear regression

$$Y = \alpha + \beta X$$

  - Intercept α and slope β specify a line approximating the data
- Multiple regression

$$Y = a + b_1 X_1 + b_2 X_2$$

  - Matrix form: **Y** = **A** + **BX**
  - Find parameters **A** and **B** by solving matrix equations

Polynomial Regression

- If data has polynomial (e.g., quadratic) form

$$Y = a + bX + cX^2 + dX^3$$

- Regard each power of X as a linear variable

$$Y = a + bX_1 + cX_2 + dX_3$$

Logarithmic regression

- Suppose data shape is exponential

$$Y = AX^{\beta}$$

- Logarithmic regression:

$$\ln Y = \alpha + \beta \ln X$$

- Multiple logarithmic regression

$$\ln Y = a + b_1 \ln X_1 + b_2 \ln X_2$$

Locally Weighted Regression

- Construct an explicit approximation to f over a local region surrounding query instance x_q.
- Locally weighted linear regression:
  - The target function f is approximated near x_q using the linear function: $\hat{f}(x) = w_0 + w_1 a_1(x) + \cdots + w_n a_n(x)$
  - minimize the squared error: distance-decreasing weight K

$$E(x_q) \equiv \frac{1}{2} \sum_{x \in k\_nearest\_neighbors\_of\_x_q} (f(x) - \hat{f}(x))^2 K(d(x_q, x))$$

  - the gradient descent training rule:

$$\Delta w_j \equiv \eta \sum_{x \in k\_nearest\_neighbors\_of\_x_q} K(d(x_q, x))((f(x) - \hat{f}(x)) a_j(x)$$

- In most cases, the target function is approximated by a constant, linear, or quadratic function.

Regression issues

- What shape does the data have?
  - Piecewise linear vs. quadratic vs. …
- How does one avoid overfitting or underfitting?
- How does one handle outliers?

Classification by regression

- Classifying data by predicting characteristic functions
- Set S has characteristic function χ_S
  - $\chi_S(d) = 1$ if $d \in S$
  - $\chi_S(d) = 0$ if $d \notin S$
- Use regression to predict χ_C for each class C in the training data
- Classify d in class with largest predicted value
  - Compare $\chi_A(d)$ with $\chi_B(d)$
  - Assign d to class A if greater, to class B otherwise
- This is called multiresponse regression

Justifying multiresponse classification

- Multiresponse classification rule seeks find a function *f* that minimizes

$$E_y\{(f(X) - Y)^2 \mid X = x\}$$

  - f(X) is the model value
  - Y is the observed target value (0 or 1)
  - x is the instance
- Algebraically equivalent to minimizing

  $(f(X) - P(Y = 1 \mid X = x))^2 + E_y\{(P(Y = 1 \mid X = x) - Y)^2 \mid X = x\}$

  $(f(X) - P(Y = 1 \mid X = x))^2 + \text{constant term}$

Pairwise regression

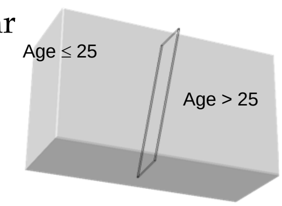
- Combines regression models with voting
- Identify a regression function for each pair of classes
  - Construct the regression function using only instances of the two classes
  - Predict output of +1 for first class, -1 for other
- To classify a new instance
  - Each pair-function "votes" for one class
  - As class that receives most votes
  - Or as "unknown" if votes not unanimous
- Usually more accurate, but more expensive than multiresponse regression

Linear classification

- Regression methods seek models linear over some set of basis functions
  - Linear basis (numeric attribute vectors)
  - Nonlinear basis (polynomials, logs, etc.)
- Multi-response linear regression separates classes with hyperplanes
  - Classify item *a* as class C_1 instead of C_2 if

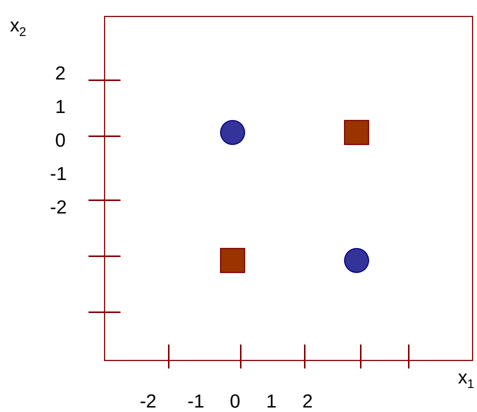
$$(w_0^{(1)} - w_0^{(2)})a_0 + \ldots + (w_n^{(1)} - w_n^{(2)})a_n > 0$$

- Similarly for pairwise linear regression



Linearity: it ain't necessarily so

The exclusive-or problem



Index *i*	*x*	*y* (class)
1	(1,1)	1
2	(1,-1)	-1
3	(-1,-1)	1
4	(-1,1)	-1

■ = class 1
● = class -1

Nonlinear classification

- Nonlinear regression allows prediction of nonlinear functions
- Nonlinear classification allows classification that does not fit linear boundaries
- Common approach
  - Transform data into new space using nonlinear mapping
  - Find linear model or boundaries
  - Return to original space by inverse mapping

Logistic regression

- Designed for classification problems
- Linear probability model for class odds ratio

$$\log(P/1\text{-}P) = w_0 a_0 + \dots + w_n a_n$$

Polynomial classification

- Polynomials form a simple nonlinear space
- All products of *n* linear attributes (degree *n*)
- Example: two attributes and 3 factors

$$w_o a_1^3 + w_o a_1^2 a_2^1 + w_o a_1^1 a_2^2 + w_o a_2^3$$

- Question: how many coefficients in a polynomial over *m* attributes of degree *n*?

Polynomial classification

- Polynomial models can be slow
  - Degree 5 model over 10 attributes has more than 2000 coefficients
  - Each coefficient constitutes an attribute for regression
  - Linear regression has time cubic in number of attributes
- Polynomial models prone to overfitting
  - Many coefficients relative to number of training examples
  - The curse of dimensionality