Regression

Adrian Barbu

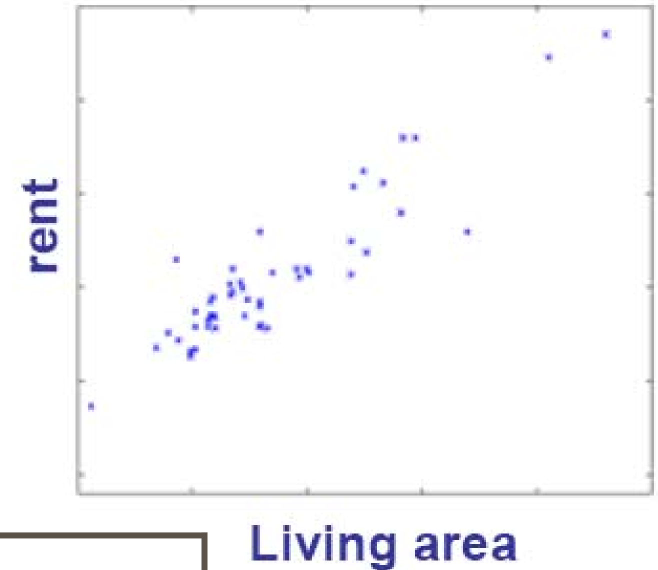
Apartment Rent Modeling

- Want to predict the rent amount
- We are given:
  - The living area (sq ft) of the apartment
  - Number of bedrooms
  - Distance from campus
- Learning=Target function approximation
- Training set:

Living Area	# Bedrooms	Dist from campus	Rent
655	1	2.5	700
550	1	2.1	600
820	2	1.8	1000
830	2	1.1	1200
…	…		…

- Test set

675	1	1.5	?
800	2	2.2	?



Problem Setup

- Features (aka predictors or attributes)
  - Area, #Bedrooms, dist from campus, …
  - Obtain a feature vector $\mathbf{x} = (x_1, ..., x_p)$
- Target
  - Rent amount y
- Training Set $\{(\mathbf{x}_i, y_i), i = 1, ..., N\}$

- Learn a function $y = f(\mathbf{x})$ that best interpolates

$$\{(\mathbf{x}_i, y_i), i = 1, ..., N\}$$

Linear Regression

- Assume f is linear in the predictors

$$f(\mathbf{x}) = f_\beta(\mathbf{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

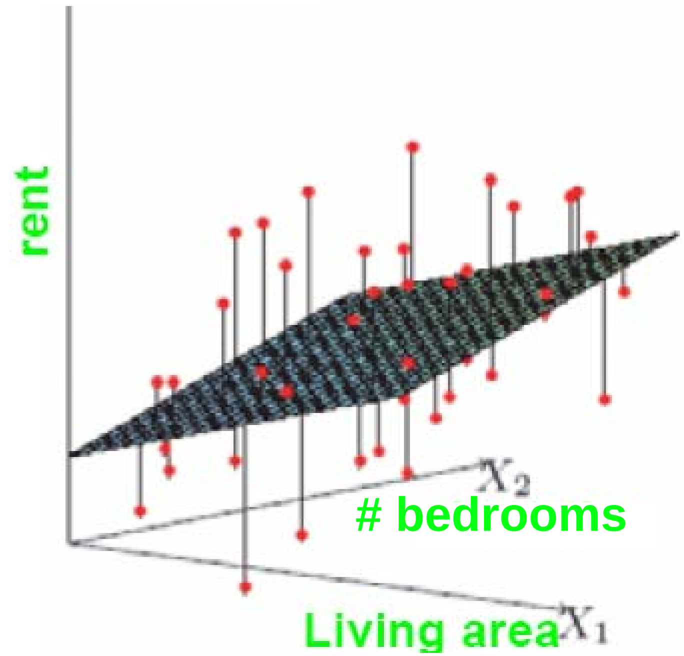
- A measure (cost) of how well it interpolates the training data

$$L(\beta) = \sum_{i=1}^{N} (y_i - f_\beta(\mathbf{x}_i))^2 + \lambda ||\beta||^2$$

- Write

$$X = \begin{pmatrix} 1 & \mathbf{x}_1 \\ ... & \\ 1 & \mathbf{x}_N \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & ... & x_{1p} \\ ... & . & . & \\ 1 & x_{N1} & ... & x_{Np} \end{pmatrix}$$

$$Y = \begin{pmatrix} y_1 \\ ... \\ y_N \end{pmatrix}$$



Linear Regression

- Then $f(X) = X\beta$
- Can write the cost in matrix form

$$L(\beta) = (Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta$$

- Quadratic function in β, one global minimum
- Set partial derivatives to zero
- By matrix calculus (or tedious computation) obtain the normal equations

$$X^T X\beta + \lambda\beta = X^T Y$$

- Obtain

$$\beta = (X^T X + \lambda I_p)^{-1} X^T Y$$

Linear Regression

- Analytical solution

$$\beta = (X^T X + \lambda I_p)^{-1} X^T Y$$

  Pros and cons:
  - Pros: Easy to implement, obtain result in one step
  - Cons: memory problems, stability is $X^T X$ is almost singular
- Gradient descent

$$\beta^{t+1} = \beta^t + \eta(X^T Y - X^T X \beta^t - \lambda \beta^t)$$

  Pros and cons:
  - Pros: Obtain a result even if $X^T X$ is singular
  - Cons: iterative, need to wait to converge

Cross-validation

- A more reliable estimate of prediction accuracy
- K-fold cross-validation:
  1. Divide all data into K subsets randomly
  2. Repeat for j=1 to K:
     1. Train an all data except subset j
     2. Evaluate prediction accuracy on subset j, obtain ε_j
  3. Final prediction accuracy is the average of all ε_j

$$\epsilon = \frac{1}{K}\sum_{j=1}^{K}\epsilon_j$$

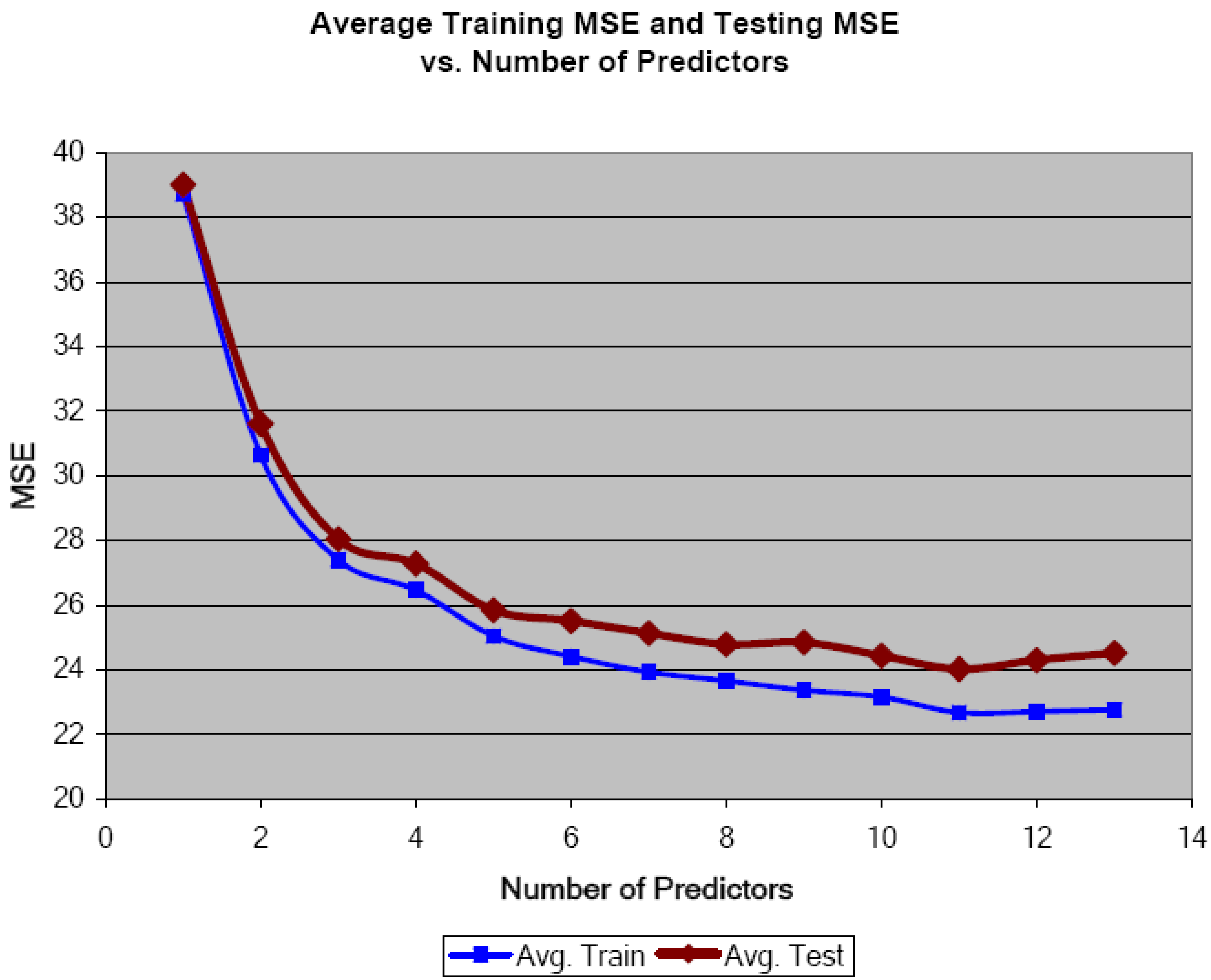
- Usually K=4-10 or even more
- In some cases K=N (the number of samples)
  - = Leave one out cross-validation

Example: Boston Housing Data

- 506 observations
- 13 predictors:
  1. crime - per capita crime rate by town
  2. zoned - proportion of residential land zoned for lots over 25,000 sq.ft.
  3. industry - proportion of non-retail business acres per town
  4. chas - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
  5. nox - nitric oxides concentration (parts per 10 million) 3.
  6. rooms - average number of rooms per dwelling
  7. age - proportion of owner-occupied units built prior to 1940
  8. dist - weighted distances to five Boston employment centres
  9. hwy - index of accessibility to radial highways
  10. tax - full-value property-tax rate per $10,000
  11. pteacher - pupil-teacher ratio by town
  12. bk - 1000(Bk 0:63)2 where Bk is the proportion of blacks by town
  13. lstat - % lower status of the population
- Output: house value

Boston Housing Data

- Optimal number of predictors: 11
- 2-fold cross-validation



Avoiding Overfitting

- Variable selection
  - Prefer fewer variables (Occam's razor)
- Two types of model (variable) selection methods:
  - Testing-based procedures
    - Forward selection
    - Backward elimination
    - Stepwise regression
  - Criterion-based procedures
    - Akaike Information Criterion, Bayes Information Criterion
    - Adjusted R^2
    - Mallow C_p statistic

P-value

- Fit the model

$$\beta = (X^T X)^{-1} X^T Y$$

- Standardize the i-th coefficient

$$t_i = \widehat{\beta}_i / se(\widehat{\beta}_i) = \widehat{\beta}_i / \sqrt{(X^T X)^{-1}_{i+1,i+1}\widehat{\sigma}}$$

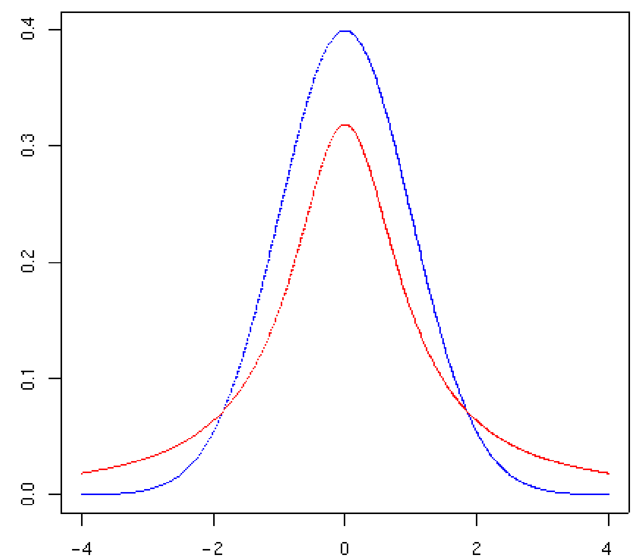
- Under the null hypothesis β_i=0:
  - t_i should follow a t-distribution (red)
- p-value = probability that for a random value ρ from the t-distribution

$$|\rho| \geq |t_i|$$

- Small p-value means β_i is significant

Red=t-distribution

Blue=Gaussian



Testing-Based Variable Selection

- Backward Elimination
  1. Fit a model with all variables
  2. Remove least significant variable with p-value at least 0.05
  3. Refit the model and repeat step 2 while possible
- Forward Selection
  1. Start with a constant model
  2. Compute the p-value for each variable when added to the model
  3. Add the most significant variable with p-value less than 0.05
  4. Repeat steps 2-3 as long as possible
- Stepwise Regression
  - A combination of Forward Selection and Backward Elimination
  - Has many variants

Criterion-Based Variable Selection

- Test all combinations of variables
- Tradeoff accuracy (RSS) with model complexity (p)
- Select best tradeoff according to a criterion:
  - Akaike Information Criterion

$$AIC = -2N\log(RSS/N) + 2p$$

  - Bayes Information Criterion

$$BIC = -2N\log(RSS/N) + p\log N$$

  - Adjusted R^2

$$R_a^2 = 1 - \frac{RSS/(N-p)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(N-1)}$$

  - Mallow C_p statistic

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - N$$

Advantages and Disadvantages

- Testing-based procedures:
  - Fast
  - Greedy, suboptimal
  - Might not solve the right problem

- Criterion-based procedures
  - Exhaustive, more powerful
  - Slow, can only deal with as many as 15-20 variables

- Sensitive to Outliers

Nonlinear Basis Functions

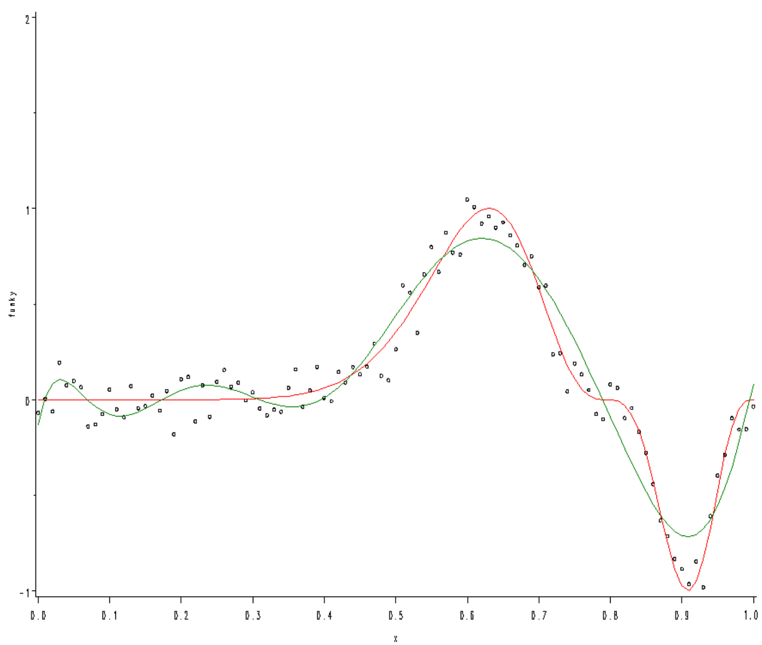
- Can transform the features through any functions we want
- Apply LR to the transformed features (basis functions)

$$y = \beta_0 + \beta_1 \phi_1(\mathbf{x}) + \ldots + \beta_k \phi_k(\mathbf{x}) = \beta^T \phi(\mathbf{x})$$

- E.g. polynomial regression

$$\phi = (1, x, x^2, \ldots, x^k)$$

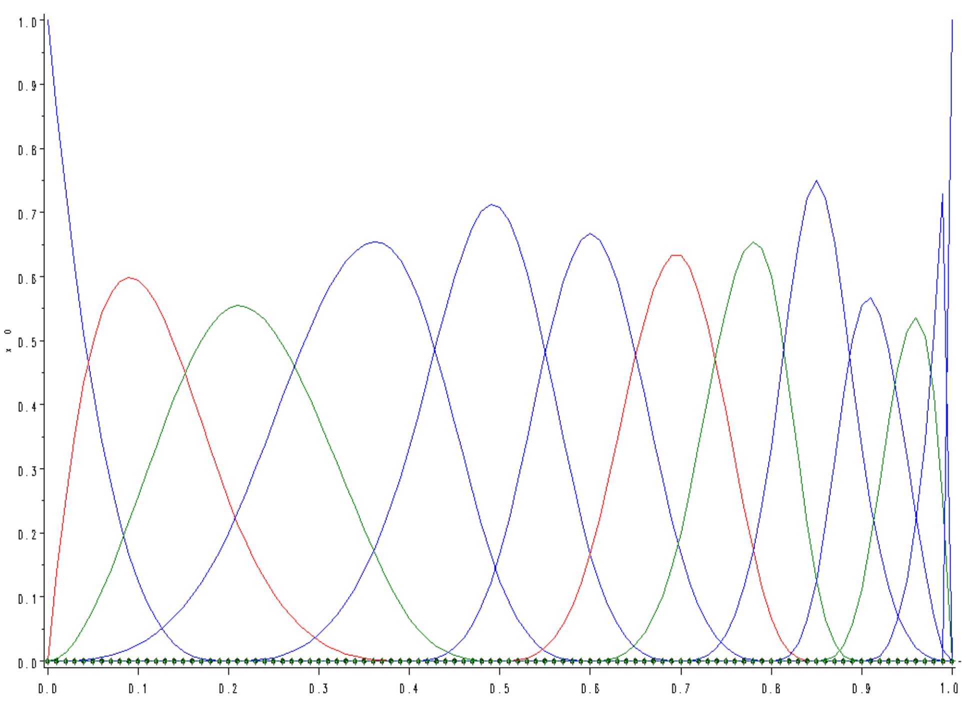
$$y = \beta_0 + \beta_1 x + \ldots + \beta_k x^k$$

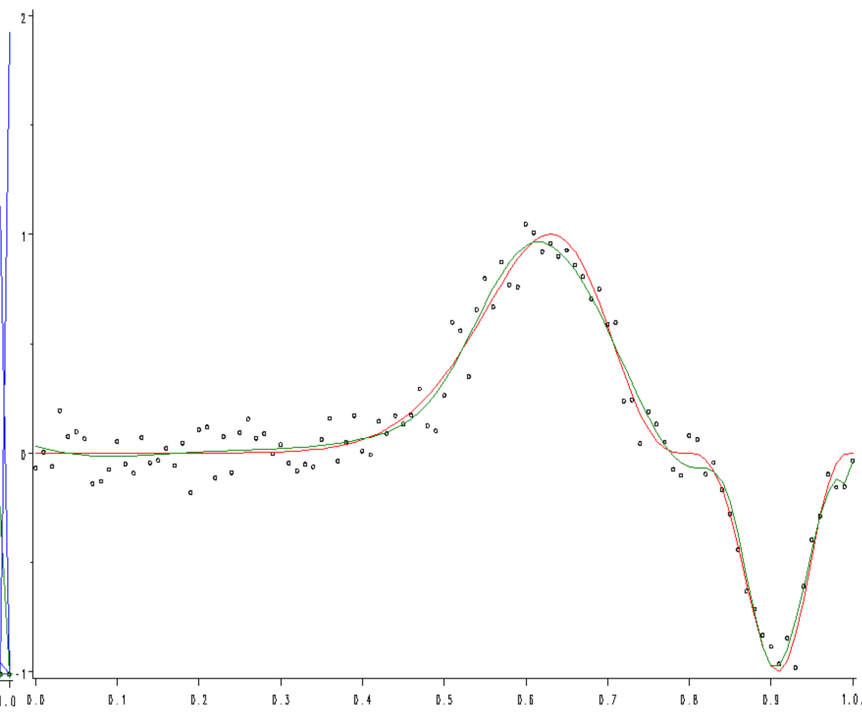


Spline Basis Functions

Basis functions

- A set of interval points $t_0, t_1, \ldots, t_k$ called knots
- Degree 3 polynomial on each interval $[t_i, t_{i+1}]$
- C^1 continuity
- Basis function i is zero outside $[t_i, t_{i+4}]$



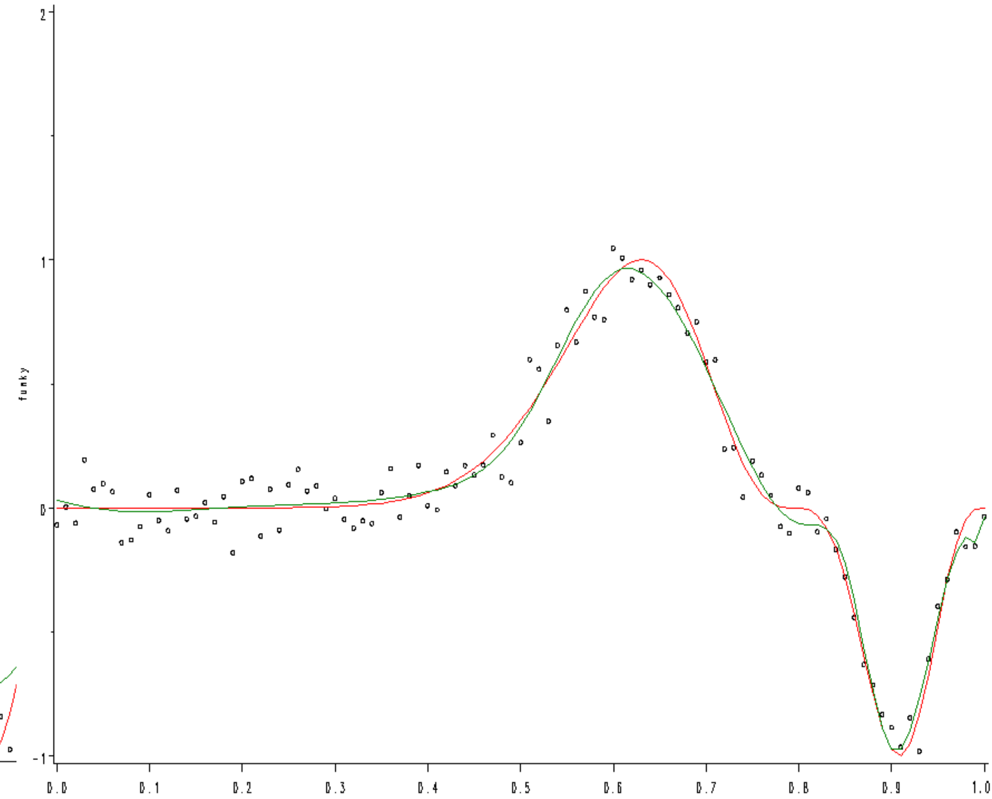


Spline vs Polynomial Basis Functions

- Splines are local and more flexible
- Polynomials have long range interactions (not good)

Polynomial (green)

Original function (red)



Spline (green)

Original function (red)

Robust Regression

- Quadratic Cost Function → sensitive to outliers
- One solution: M-estimation
  - Minimize

$$Cost(\beta) = \sum_{i=1}^{N} \rho(y_i - \mathbf{x}_i\beta)$$

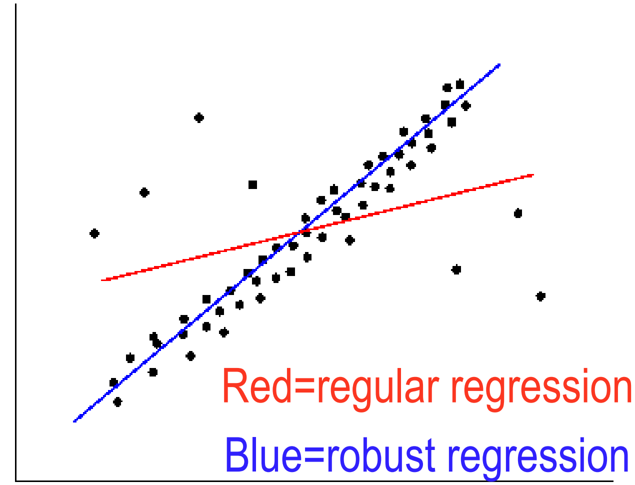
  - Regular regression has $\rho(x) = x^2$
  - Least Absolute Deviation (LAD)

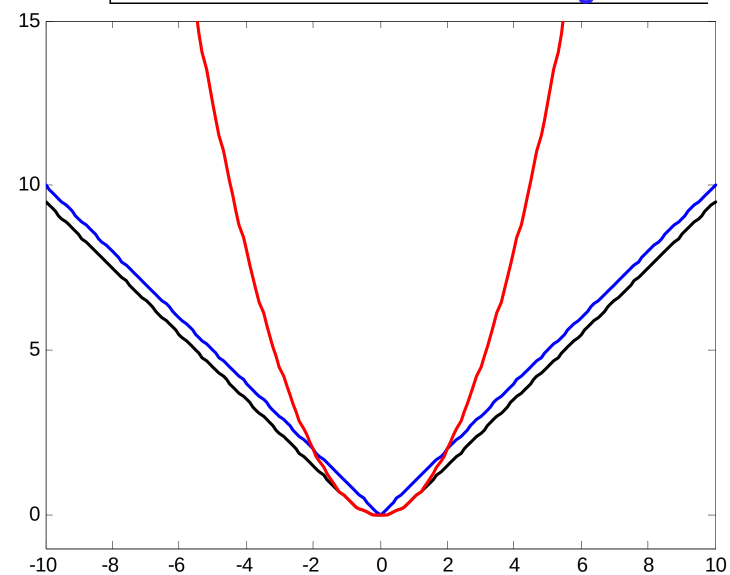
$$\rho(x) = |x|$$

  - Huber

$$\rho(x) = \begin{cases} x^2/2 & \text{if } |x| < c \\ c|x| - c^2/2 & \text{otherwise} \end{cases}$$

ρ gives smaller weight to large errors





Functions ρ

M-Estimation

- M-estimation=Convex optimization
- One algorithm:

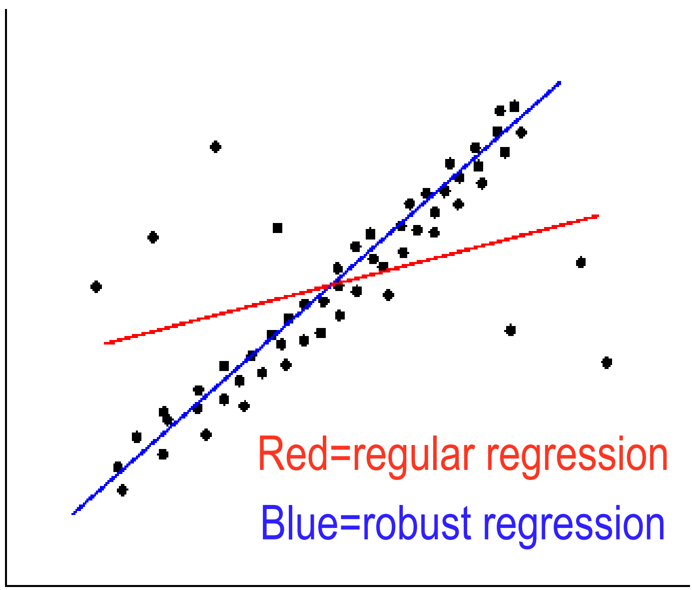
1. Initialize β
2. Compute weights

$$w_i = \rho'(y_i - \mathbf{x}_i\beta)/(y_i - \mathbf{x}_i\beta)$$

3. Reestimate β by weighted regression

$$Cost_W(\beta) = \sum_{i=1}^{N} w_i(y_i - \mathbf{x}_i\beta)^2$$

4. Repeat 2-3 until convergence



Robust Regression

- Another solution: Least Trimmed Squares
  - Robust cost function

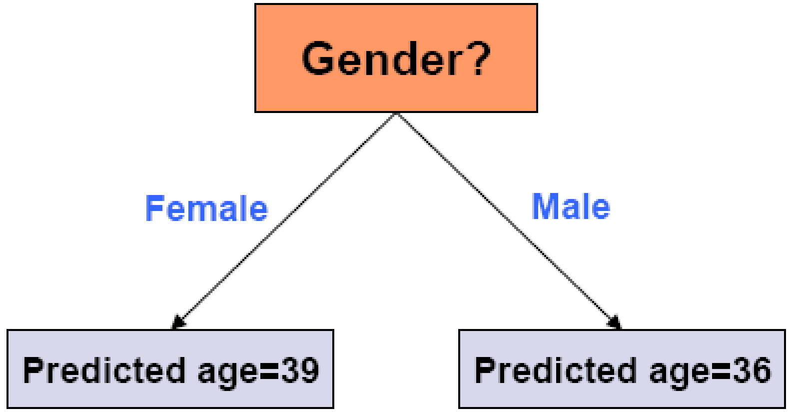
$$Cost(\beta) = \sum_{i=1}^{q} \hat{\epsilon}_{(i)}^2$$

    where $\hat{\epsilon}_{(i)}$ are the sorted residuals $\hat{\epsilon}_i = y_i - \mathbf{x}_i\beta$ (in increasing order) and q is about N/2
  - Largest errors do not influence the cost (hence robustness)
  - Minimization is tricky since Cost is not convex
    - Use a random method such as MCMC
  - Slower but more powerful than M-estimation

Regression Tree

- Decision tree for regression

Gender	Rich?	Num. Children	# travel per yr.	Age
F	No	2	5	38
M	No	0	2	25
M	Yes	1	0	72
:	:	:	:	:



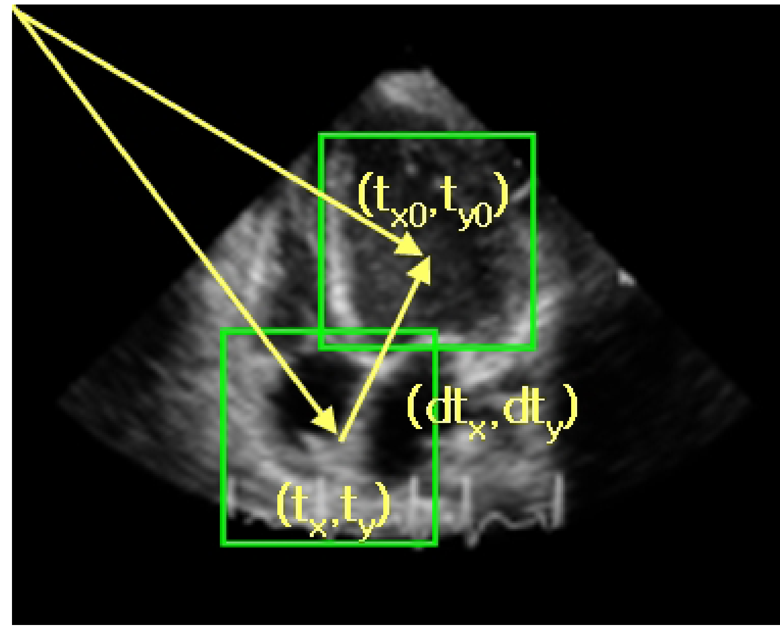
Object Detection and Context

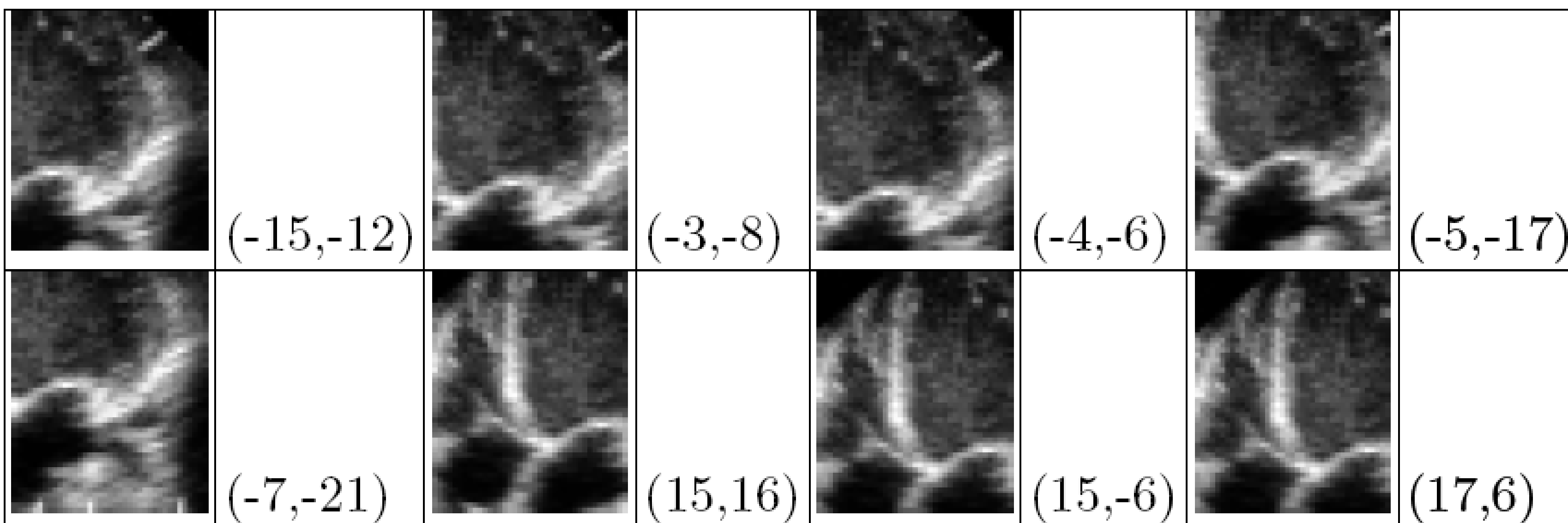

Application: Image Based Regression

- Given a location, want the relative position of the LV Center

Training

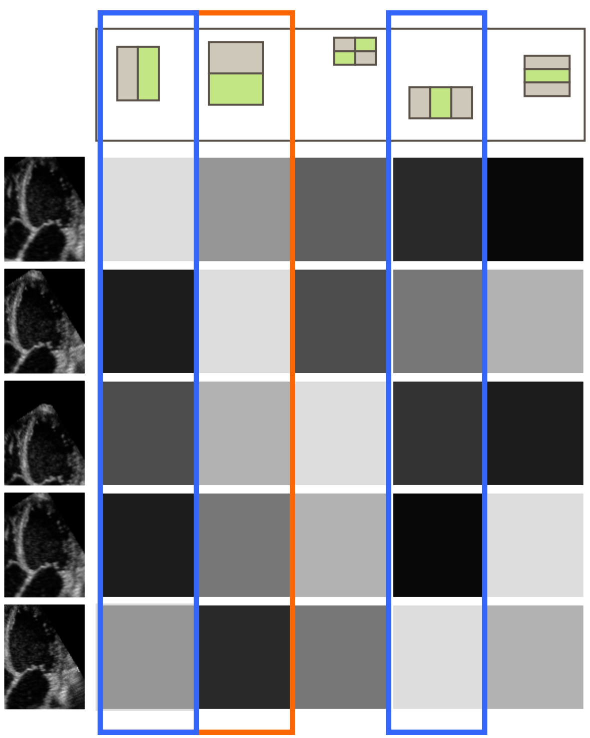
- Each coordinate is trained independently
- Training data: 13,500 pairs
  image + relative location





Regression Details

- 10,000 predictors (Haar features)
- Only 200 are selected to be used
- Train a regressor for each coordinate independently
- Use a linear combination of locally constant regressors
  - Select best one
  - Select next one that minimizes the residual error
  - And so on



(3.2, -15.5)
(-6.2, -5.3)
(4.7, 16.5)
(-0.3, -8.7)
(-13.2, -18.0)

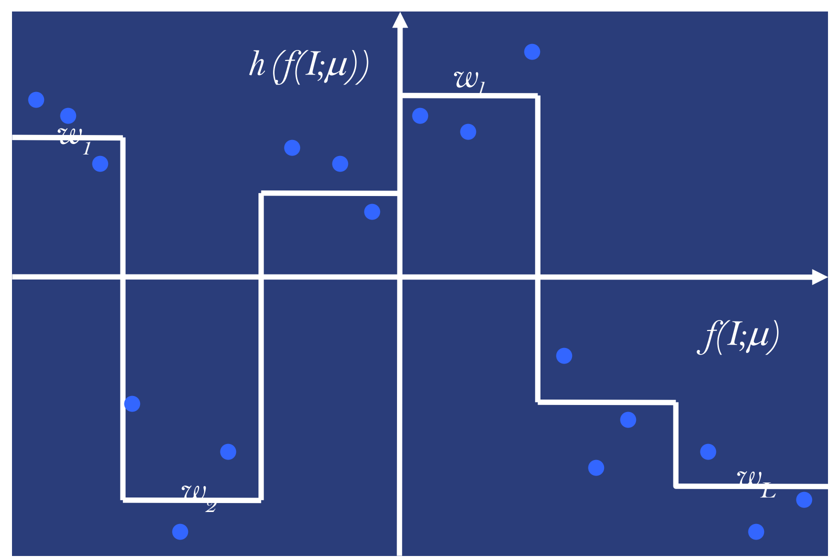
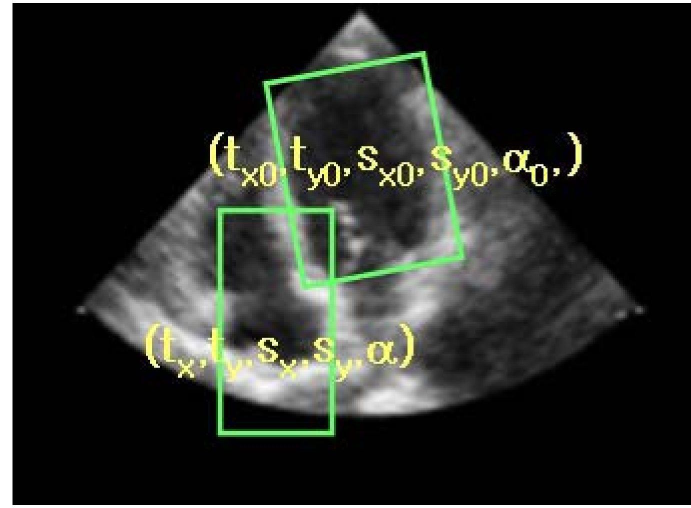
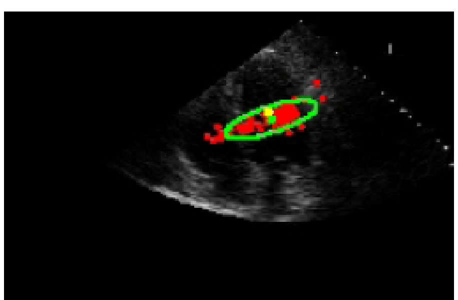


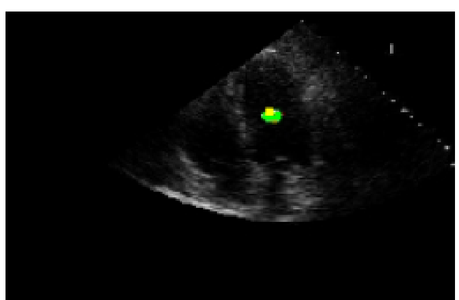
Image Based Regression

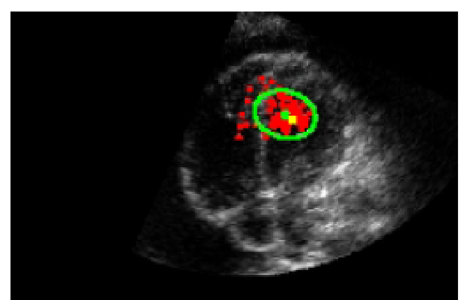
- 5 LV parameters:
  - Position (t_x,t_y)
  - Scale (s_x, s_y)
  - Rotation angle α
- Start from 200 random locations



- Use a detector based on Boosting to filter the final result
- S. Zhou and D. Comaniciu. Shape Regression Machine. Information Processing in Medical Imaging (IPMI), 2007


Top 100 locations


After detector


Top 100 locations


After detector

Conclusions

- Regression = learning target functions with continuous output

- Issues with regression:
  - Overfitting → variable selection
  - Outliers → robust regression
  - Linear vs. non-linear → polynomial, splines, locally constant

- Applications:
  - Apartment rent, house prices
  - Diabetes, prostate, BMI, age
  - Relative location of LV