Data Mining & Machine Learning

CS37300
Purdue University

Sep 29, 2023

Linear Regression

Setup

- Data $\{x_i, y_i\}$ Total *n* data points, $x_i \in R^d, y \in R$.
- $y \sim w^\top x$
- Cost function:
  - $\min_w \sum_i |y_i \ - w^\top x_i|^2$
- Interpretation: Minimize the ”residuals” $(y_i - w^\top x_i)$ i.e. leftover after fitting the model

Closed form solution

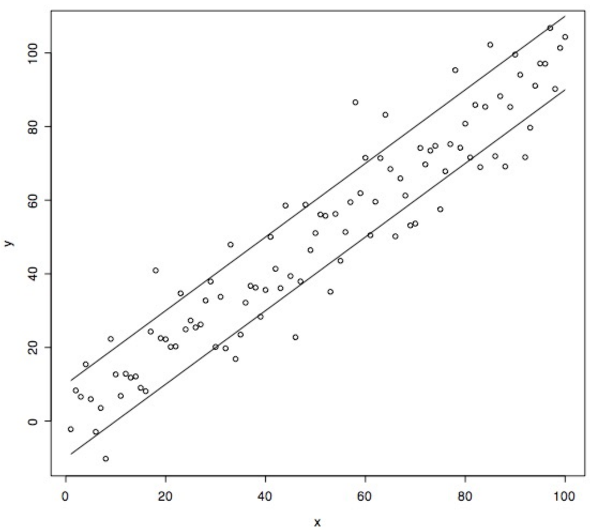
- $w^{\star} = (X^T X)^{-1} X^T y$
- Prediction: $w^{\star T} x_{test}$
- What is the time and space complexity ?
  - O(d*d*n + d*d*d)
- What if $X^{\top} X$ is not invertible? When does this happen?
  - Multicollinearity in the features (columns)
  - Instability: Small change in input can alter the output a lot. Makes interpretation hard

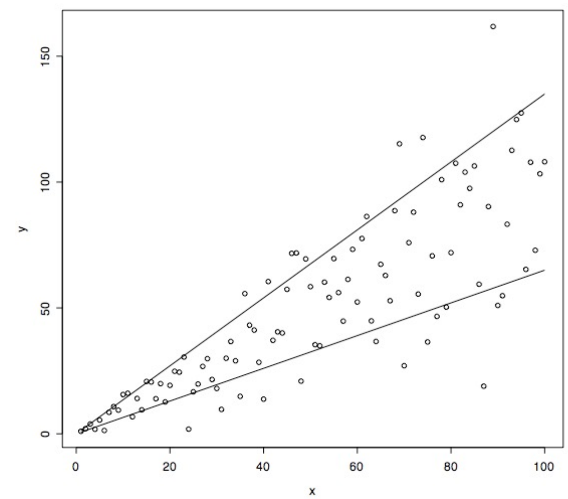
Eigendecomposition

- Let A be a square matrix with N linearly independent eigenvectors, q_i (i=1, …, N). Then A can be factorized as:
  - $A = Q\Lambda Q^{-1}$
  - Q is the square matrix whose i-th column is the eigenvector q_i of A, Λ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, i.e., $\Lambda_{ii} = \lambda_i$
  - For a symmetric matrix A, Q is an orthogonal matrix, that is, $A = Q\Lambda Q^T$
  - With Eigendecomposition available, how can we invert A ?

Probabilistic interpretation

- $y = x^{\top} w + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. What is p(y|x) ?
- MLE in this model is equivalent to linear regression,
- Clearly delineates the underlying ”implicit” assumptions in the least squared loss e.g. Homoskedasticity. Also, quantifies uncertainty.





- Probabilistic interpretation also quantifies uncertainty under the same assumptions

Generalized Linear Models (GLMs)

- Y is a RV from a distribution in the exponential family e.g. Gaussian, Bernoulli, Poisson etc
- For some function $g(.)$
  - $g(E[Y]) = (x^T w)$
- Allows for certain algorithms to be applied widely to all GLMs.

A taste of interpretability

- How important is each individual feature ?
- Recall we can write: $y_j = \sum_i w_i x_{ji}$
- A measure of feature importance: "If x_{ji} is increased by 1, how much does y_j change?
- Sizes of individual w_i tell us about relative importance of features
  - Counter-example: Should income and credit score be on the same scale?

A taste of evaluation

- How do we check whether the learnt model is reasonable?
- MSE
- Coefficient of determination or R^2 statistic = $\frac{\sum_i (y_i - w^T x_i)^\wedge 2.}{\sum_i (y_i - \bar{y})^\wedge 2}$
  - Intuitively, ratio of variance explained and total variance

Ridge regression

- Alternate cost function using regularization:
  - $\min_{w} \sum_i |y_i - w^\top x_i|^2 + \lambda ||w||^2$
- "Want to fit the data but don't want w to be too big"
- Also has closed form solution: $w^\star = (X^T X + \lambda I)^{-1} X^T y$
  - The inner matrix is always invertible
- Bias-variance tradeoff (more on this later!)

Sparse modeling

- Instead of using all the features, start with using a few
- Iteratively add or remove features that improve "performance" (e.g. MSE or R^2)
- Alternatively, use LASSO

$$\min_w \sum_i |y_i - w^\top x_i|^2 + \lambda ||w||_1 \quad ,$$

where $||w||_1 = \sum_i |w|$ is the L1 norm of w.

- Commonly used: Elastic net

$$\min_w \sum_i |y_i - w^\top x_i|^2 + \lambda_1 ||w||^2 + \lambda_2 ||w||_1$$