Regression is Everywhere

"Group 13" are: Yi Tong and Dick G1

May 3, 2016

Overview

Type of Dependent Variables				Parameter Estimate	Regularization
Continues	Linear Regression			OLS LAD	Ridge Lasso Elastic Net
Categorical	Logistic Regression	Multinomial	Ordinal		
	Probit Regression	Multinomial	Ordinal		
Count Data	Poisson Regression				

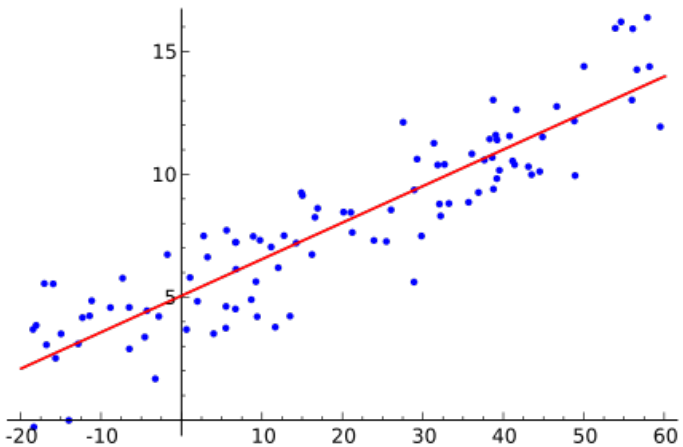
Linear Regression

- Linear Function of parameters w_j
- deterministic function

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D$$

- with additive Gaussian noise

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon$$



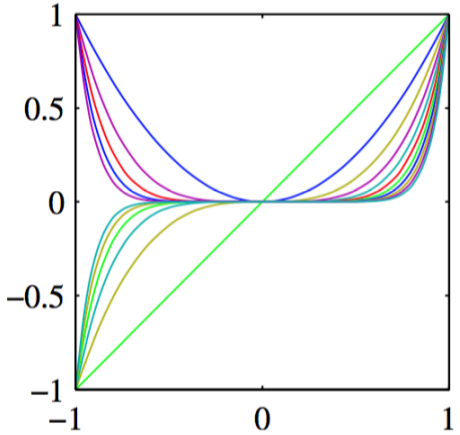
wiki

Linear Regression

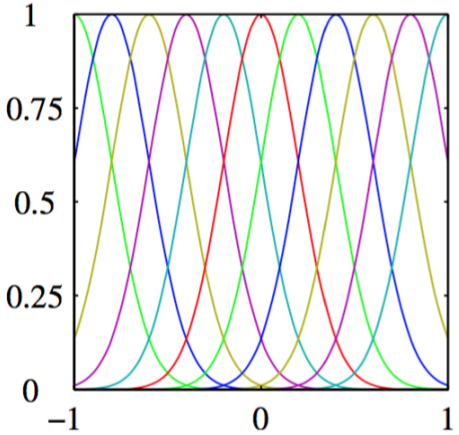
- More General:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

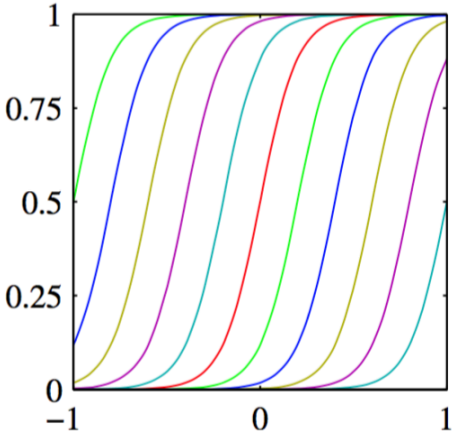
Basic Function $\phi_j(\mathbf{x})$



Polynomials



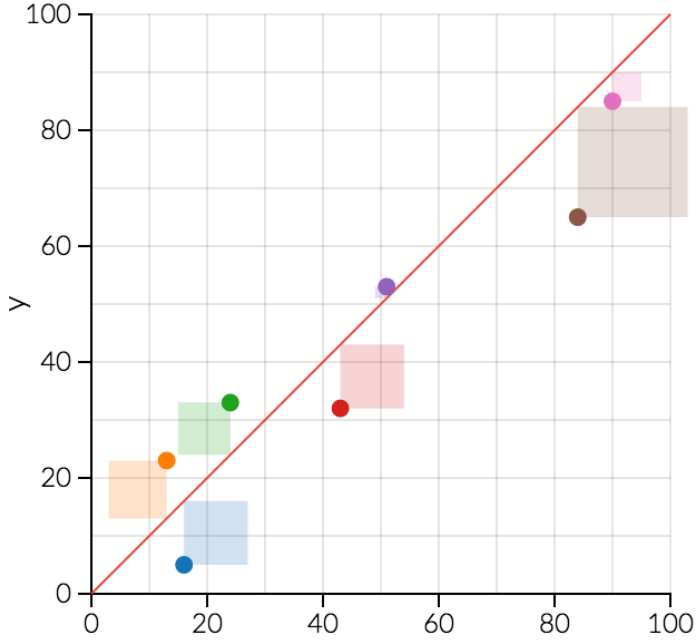
Gaussians

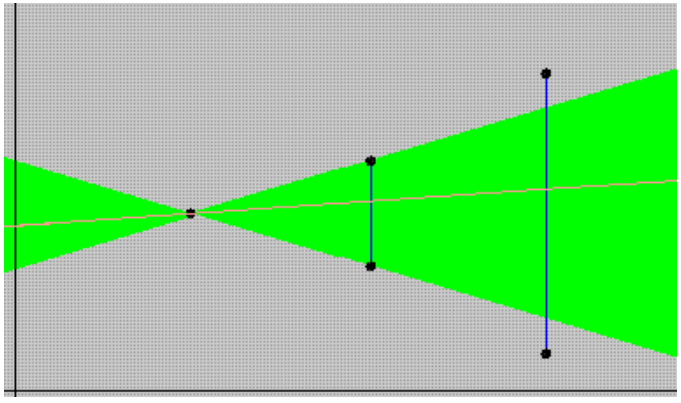


Sigmoidal

Pattern Recognition and Machine Learning Figure 3.1

OLS Ordinary Least Squares and LAD Least Absolute Deviations





Advantage
- resistant to outliers
- robust to departures from the normality assumption

Disadvantage
- computationally expensive than OLS
- having the possibility of more than one solution

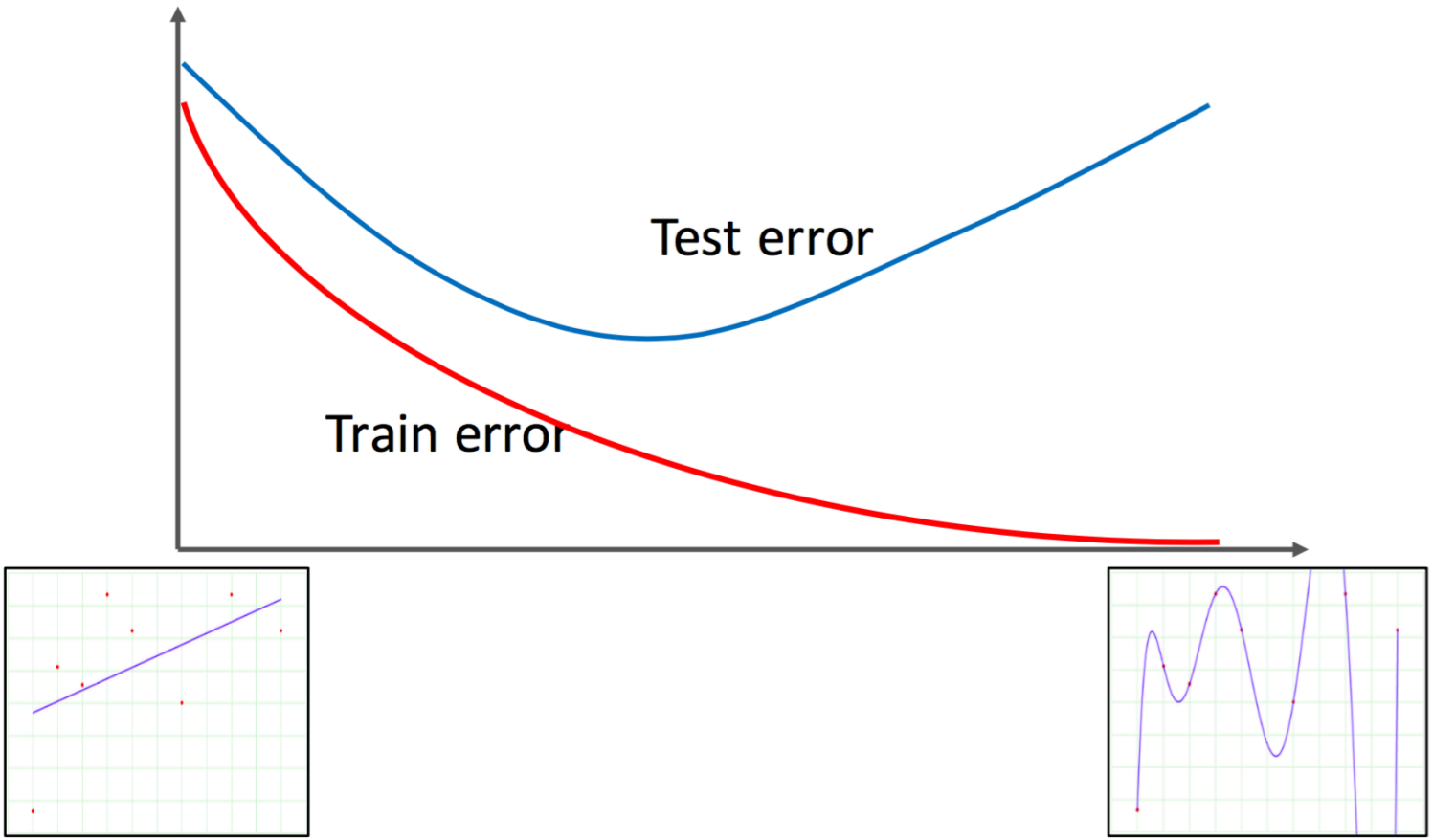
$$E_{OLS} = \sum (y_i - f(x_i))^2$$

$$E_{LAD} = \sum |y_i - f(x_i)|$$

http://setosa.io/ev/ordinary-least-squares-regression/

wiki

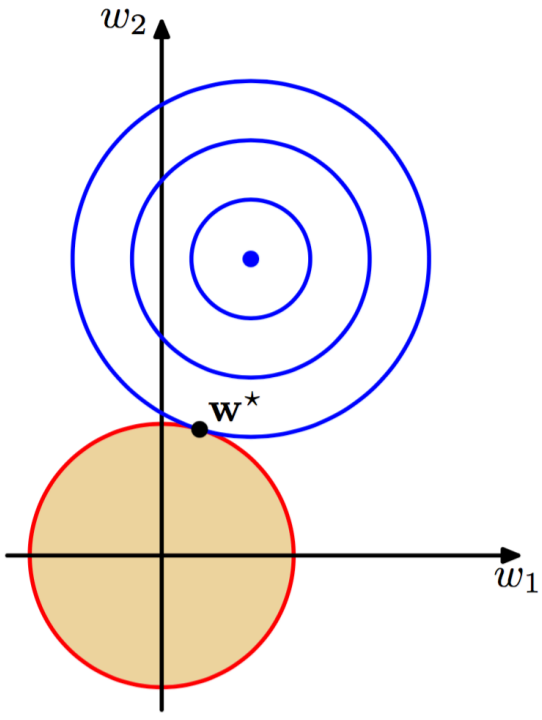
Overfitting

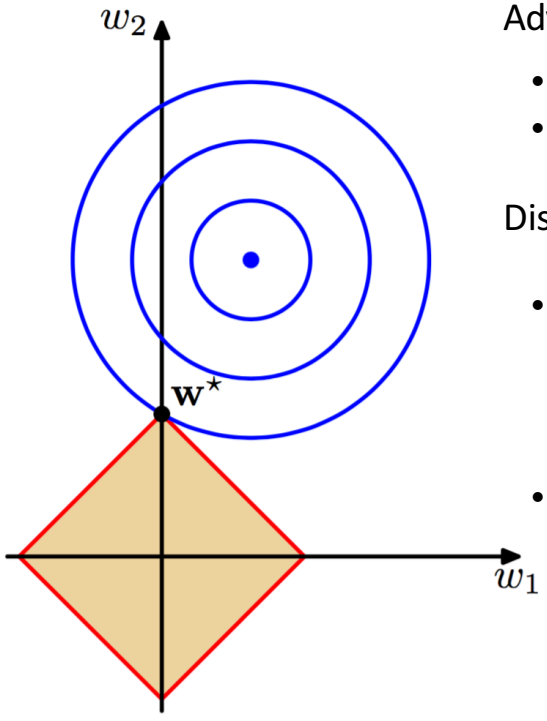


Cross Validation

- Covered in class

Ridge Regression and Lasso





Advantage

- parameter shrinkage
- variable selection

Disadvantage

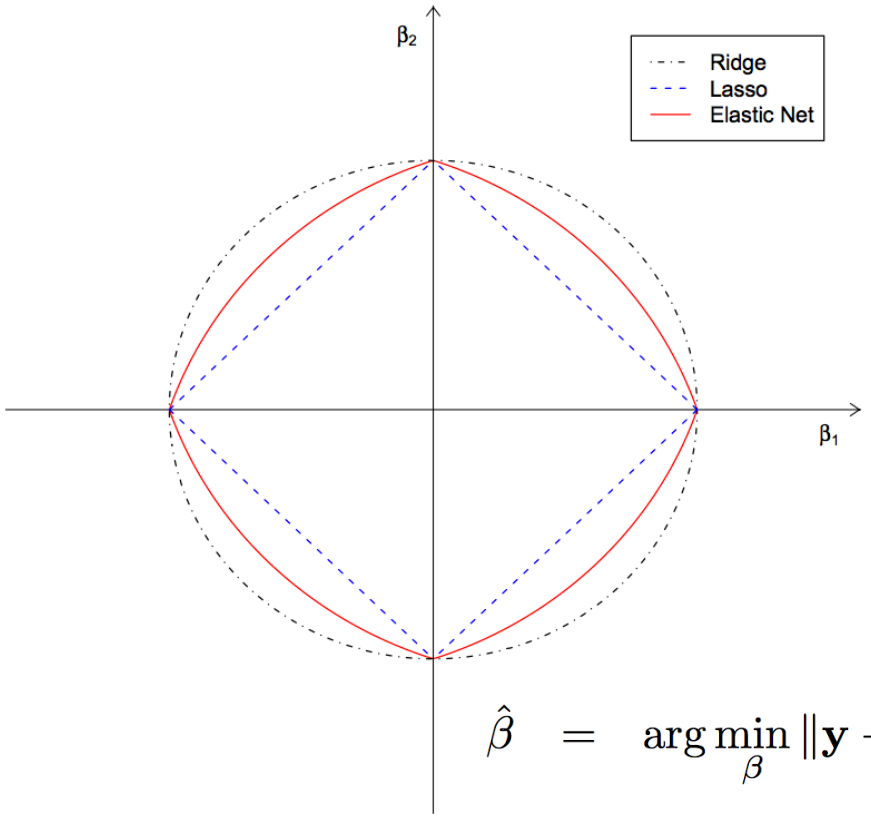
- If predictor variable are highly correlated to each other, it will select only one
- Can't do grouped selection

$$\sum (w_j)^2 < \lambda$$

$$\sum |w_j| < \lambda$$

Pattern Recognition and Machine Learning Figure 3.4

Elastic Net



- Removes the limitation on the number of selected variables;
- Encourages grouping effect;
- Stabilizes the L1 (Lasso) regularization path.

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1$$

Regularization and Variable Selection via the Elastic Net

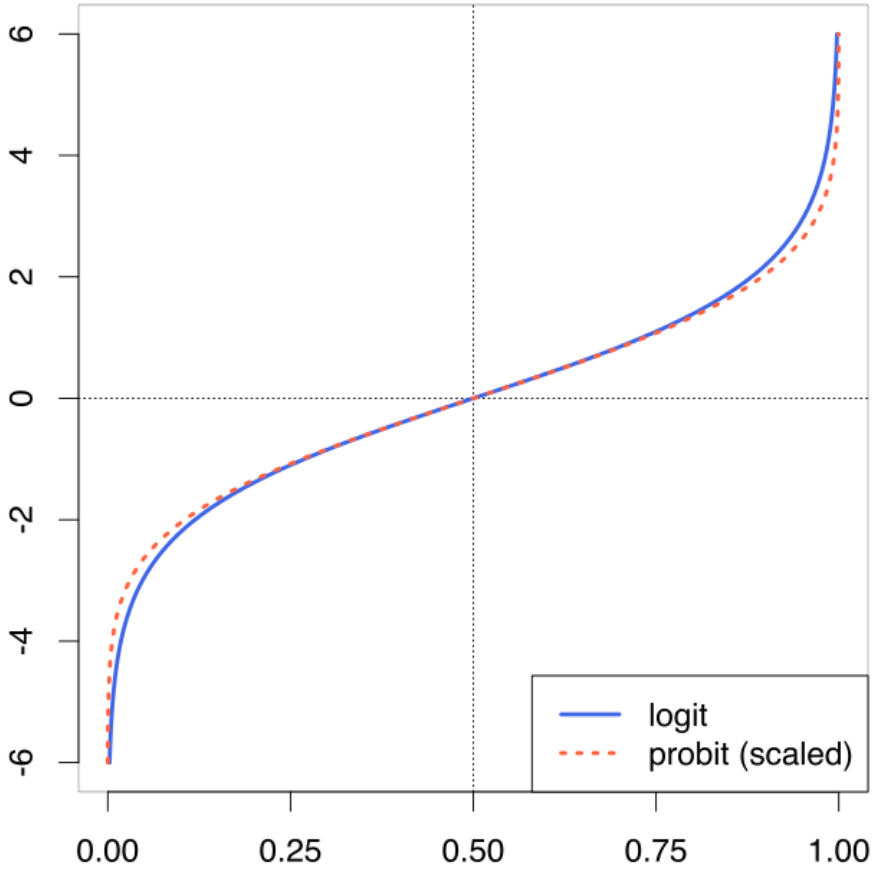
Logistic Regression

- Dependent Variable takes binary value

$$\ln\left(\frac{p}{1-p}\right) = X\boldsymbol{\beta}$$

$$P(Y = 1|X) = \frac{1}{1 + e^{-X^{-1}\boldsymbol{\beta}}}$$

Probit Regression



Logistic Regression:

$$P(Y = 1|X) = \frac{1}{1 + e^{-X^{-1}\boldsymbol{\beta}}}$$

Probit Regression

$$P(Y = 1|X) = \phi(X^{-1}\boldsymbol{\beta})$$

Where ϕ is the Cumulative Distribution Function (CDF) of the standard normal distribution

- probit curve approaches the axes more quickly

Generalized Linear Models

- Random Component: the probability distribution of the response variable (Y)
- (linear combination of) Systematic Component: $\beta_0 + \beta_1 x_1 + \beta_2 x_2$
- Link Function, *η or g(μ):* link between random and systematic components

Simple Linear Regression	Logistic Regression	Log-linear Model (special: Poisson Regression)
$e_i \sim N(0, \sigma^2)$	$Binomial(n,\pi)$	$Y_i \sim \text{Poisson}(\lambda_i)$
$\beta_0 + \beta x_i$	$\beta_0 + \beta x_i + \ldots + \beta_0 + \beta x_k$	$\lambda + \lambda_i^{X_1} + \lambda_j^{X_2} + \ldots + \lambda_k^{X_k} + \ldots$
Identity Link	Logic Link	Log Link
$\eta = g(E(Y_i)) = E(Y_i)$	$\eta = \text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$	$\eta = log(\mu)$
$E(Y_i) = \beta_0 + \beta x_i$	$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) =$ $\beta_0 + \beta x_i + \ldots + \beta_0 + \beta x_{k'}$	$\log(\mu_{ij}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$

Generalized Linear Models

Model	Random	Link	Systematic
Linear Regression	Normal	Identity	Continuous
ANOVA	Normal	Identity	Categorical
ANCOVA	Normal	Identity	Mixed
Logistic Regression	Binomial	Logit	Mixed
Loglinear	Poisson	Log	Categorical
Poisson Regression	Poisson	Log	Mixed
Multinomial response	Multinomial	Generalized Logit	Mixed

https://onlinecourses.science.psu.edu/stat504/node/216

$$A = \left(\begin{array}{c|cccccc} & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 \\ \hline \text{cosmonaut} & 1 & 0 & 1 & 0 & 0 & 0 \\ \text{astronaut} & 0 & 1 & 0 & 0 & 0 & 0 \\ \text{moon} & 1 & 1 & 0 & 0 & 0 & 0 \\ \text{car} & 1 & 0 & 0 & 1 & 1 & 0 \\ \text{truck} & 0 & 0 & 0 & 1 & 0 & 1 \end{array}\right)$$

$$\hat{A} = \left(\begin{array}{c|cccccc} & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 \\ \hline \text{Dim 1} & -1.62 & -0.60 & -0.04 & -0.97 & -0.71 & -0.26 \\ \text{Dim 2} & -0.46 & -0.84 & -0.30 & 1.00 & 0.35 & 0.65 \end{array}\right)$$

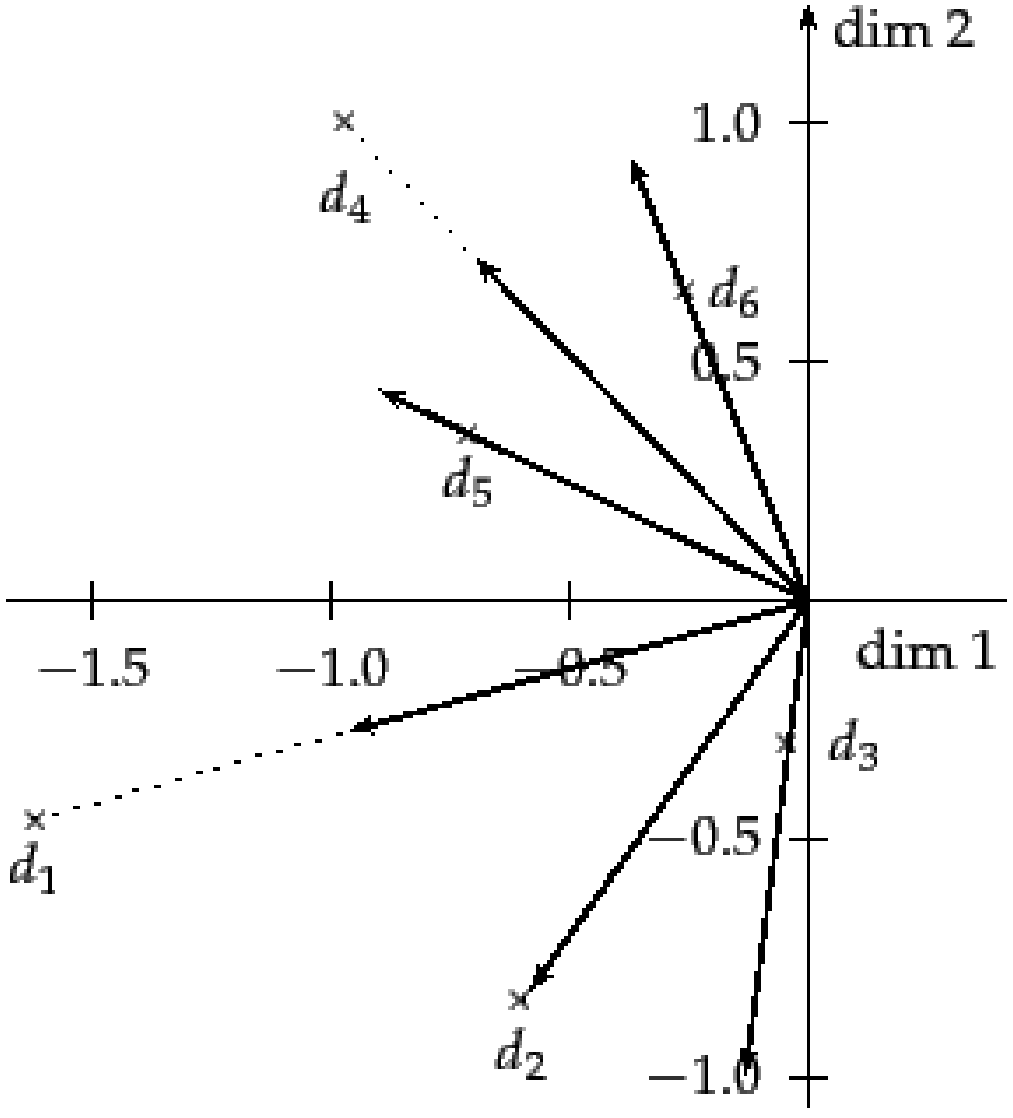


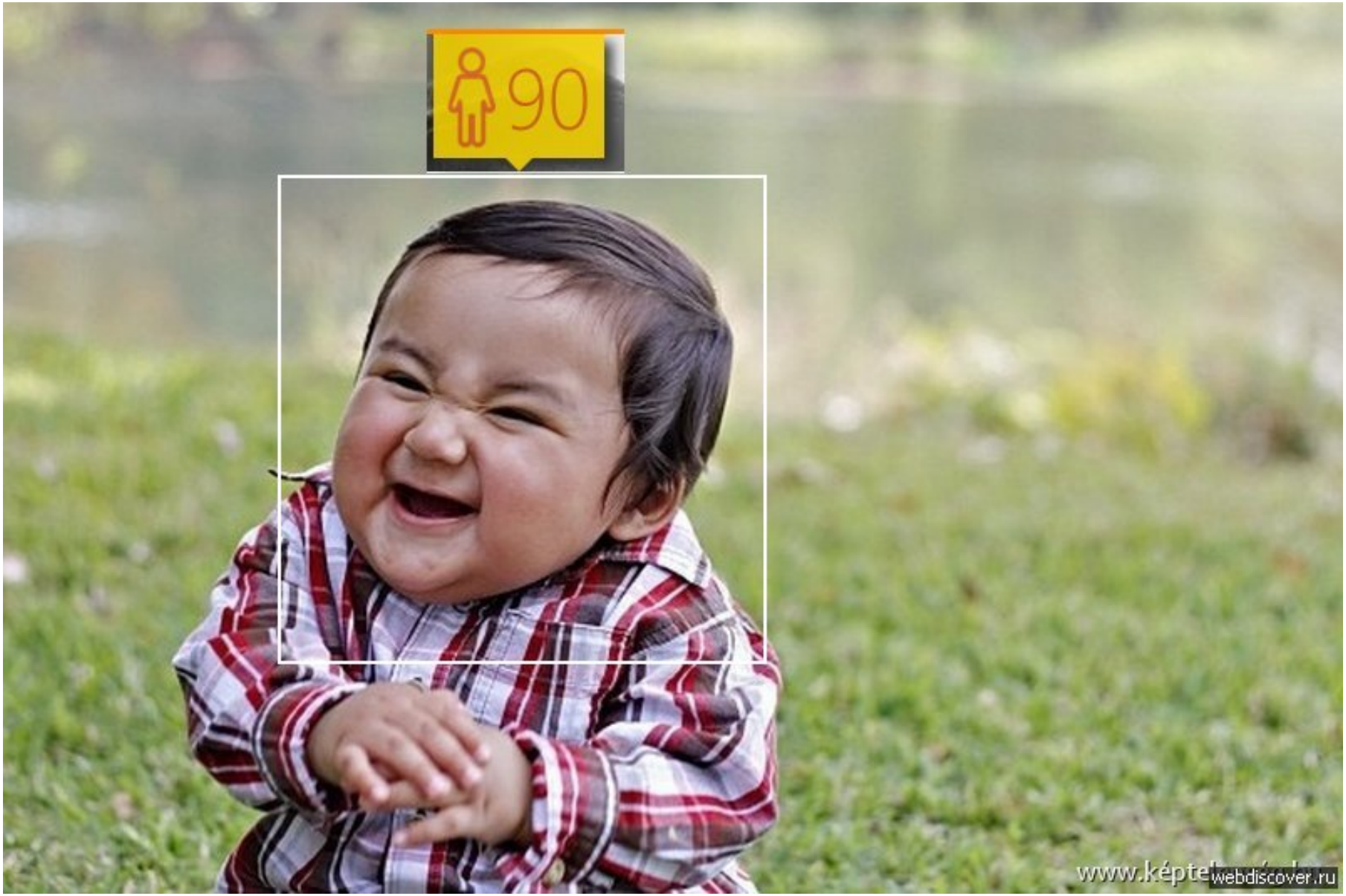
Figure : Dim 2 separates Space from Autos [1]

TABLE 3.1. WEIGHT-RATE ANALYSIS: WORDS, WEIGHTS, AND IMPORTANCES (TIMES 10^4)

	Weight	Importance		Weight	Importance		Weight	Importance
Group 1			*Group 3*			*Group 5*		
upon	1394	3847	as	−0140	0339	innovation	−1681	0336
			at	0247	0318	language	−1448	0304
Group 2			by	−0146	0542	vigor	2174	0543
although	−1754	0351	of	0037	0281	voice	−2159	0410
commonly	1333	0267	on	−0271	0706			
consequently	−1311	0459	there	0463	0972	*Group 6*		
considerable	0784	0251				destruction	1709	0342
enough	0683	0403	*Group 4*					
while	2708	0704	would	0085	0428			
whilst	−2206	0993						

Figure : Hamilton or Madison? Logistic Regression.[3]

Regression is Age Prediction



Nguyen et al, 2011 [4]

> *In order to strive for a model with high explanatory value, we use a linear regression model with Lasso regularization. This minimizes the sum of squared errors, but in addition adds a penalty term.*

actually	-.457
you	-.387
mean	-.343
definitely	-.273
pretty	-.137
huge	-.126

(a) Young

job	.514
kids	.228
years	.178
meds	.112
weekend	.094
hline	

(b) Old

Summary

- Eternally $Ax = b$, where A is (samples $\times$ features) data array, x are weights, y are labels.
- Regression on continuous and discrete y. GLMs encompass both.
- AI research moving away from formal to fuzzy logics (Markov Logic Networks).


Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999. ISBN: 0-262-13360-1.


Microsoft Azure ML. URL: how-old.net.


F. Mosteller and D.L. Wallace. *Inference and disputed authorship: The Federalist*. Addison-Wesley series in behavioral science. Addison-Wesley, 1964. URL: https://books.google.com/books?id=KKKFAAAAMAAJ.


Dong Nguyen, Noah Smith, and Rosé Carolyn. "Author Age Prediction From Text using Linear Regression". In: *Association for Computational Linguistics (ACL)*. 2011.