



Linear Regression Models

Paul Prasse, Niels Landwehr, Tobias Scheffer

Overview

- Linear regression models
- Loss functions and regularizers for regression
- Empirical risk minimization
- Special cases:
 - ◆ Lasso
 - ◆ Ridge regression
- Analytic solution for ridge regression

Regression

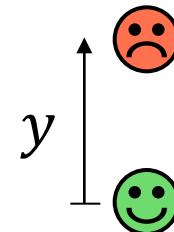
- Input: Instance $\mathbf{x} \in X$.
 - ◆ e.g., feature vector

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$$



How toxic is a combination?

- Output: continuous (real) value, $y \in \mathbb{R}$
 - ◆ e.g., *toxicity*.



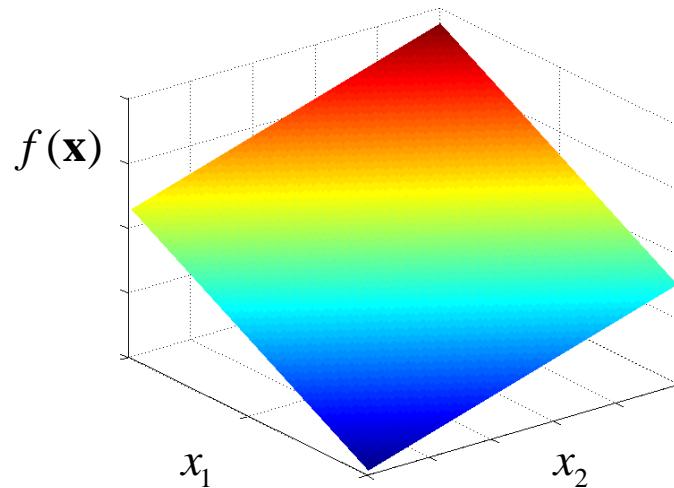
Linear Regression Models

- Regression function:

$$f_{\theta}(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\theta} + \theta_0$$

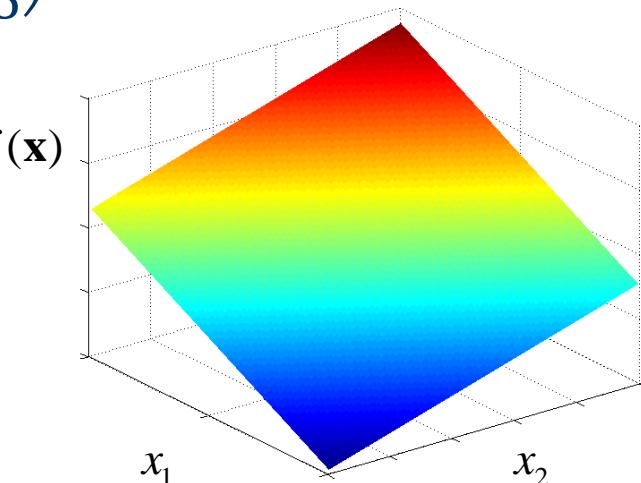
- Example:

$$f_{\theta}(\mathbf{x}) = \mathbf{x}^T \begin{pmatrix} -1 \\ 0.25 \end{pmatrix} - 2$$



Linear Regression Models

- Offset can “disappear” into parameter vector.
- Example
 - ◆ Before: $f_{\theta}(\mathbf{x}) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} -1 \\ 0.25 \end{pmatrix} - 2$
 - ◆ After: $f_{\theta}(\mathbf{x}) = \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} -2 \\ -1 \\ -0.25 \end{pmatrix}$
- New constant attribute $x_0 = 1$ added to all instances.
- Offset θ_0 integrated into θ .



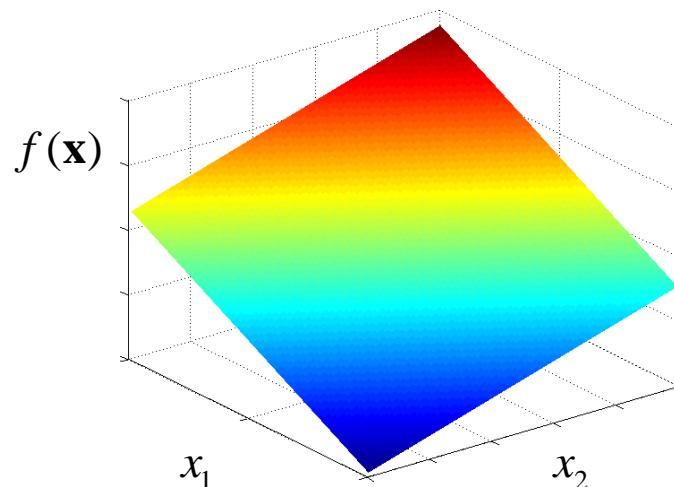
Linear Regression Models

- Regression function:

$$f_{\theta}(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\theta}$$

- Example:

$$f_{\theta}(\mathbf{x}) = \mathbf{x}^T \begin{pmatrix} -2 \\ -1 \\ -0.25 \end{pmatrix}$$



Learning Regression Models

- Input to the Learner:
Training data T_n .

- ◆ $\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}$ 

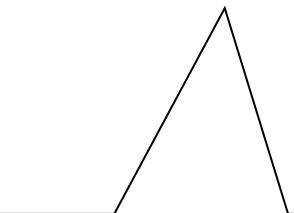
- ◆ $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ 

- Training Data:
 $T_n = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$

- Output: a model

- ◆  \mapsto 
 $y_{\theta} : X \rightarrow Y$

- ◆ For example
 $y_{\theta}(\mathbf{x}) = \mathbf{x}^T \theta$



Linear regression model with parameter vector θ .

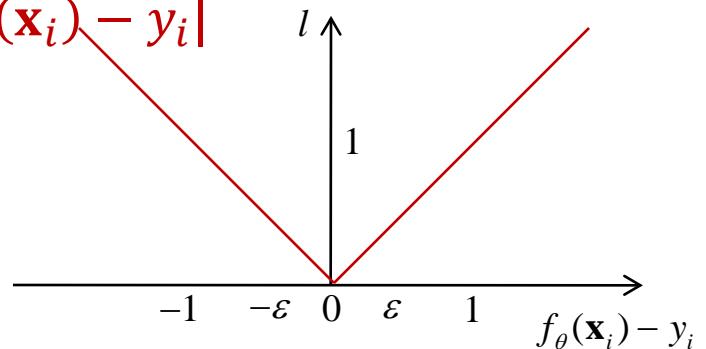
Overview

- Linear regression models
- Loss functions and regularizers for regression
- Empirical risk minimization
- Special cases:
 - ◆ Lasso
 - ◆ Ridge regression
- Analytic solution for ridge regression

Loss Functions for Regression

- Absolute loss:

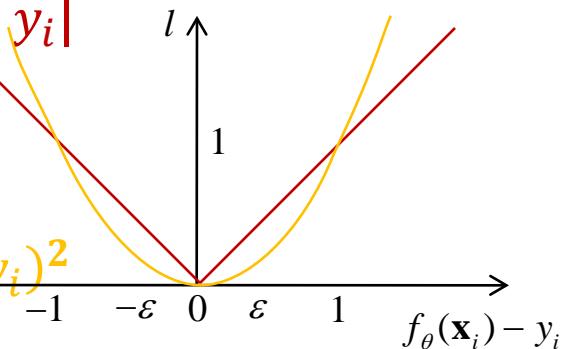
$$\ell_{\text{abs}}(f_{\theta}(\mathbf{x}_i), y_i) = |f_{\theta}(\mathbf{x}_i) - y_i|$$



Loss Functions for Regression

- Absolute loss:

$$\ell_{\text{abs}}(f_{\theta}(\mathbf{x}_i), y_i) = |f_{\theta}(\mathbf{x}_i) - y_i|$$



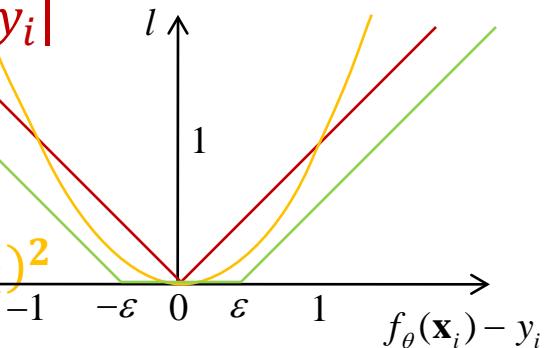
- Squared loss:

$$\ell_2(f_{\theta}(\mathbf{x}_i), y_i) = (f_{\theta}(\mathbf{x}_i) - y_i)^2$$

Loss Functions for Regression

- Absolute loss:

$$\ell_{\text{abs}}(f_{\theta}(\mathbf{x}_i), y_i) = |f_{\theta}(\mathbf{x}_i) - y_i|$$



- Squared loss:

$$\ell_2(f_{\theta}(\mathbf{x}_i), y_i) = (f_{\theta}(\mathbf{x}_i) - y_i)^2$$

- ε -insensitive loss:

$$\ell_{\varepsilon}(f_{\theta}(\mathbf{x}_i), y_i) = \begin{cases} |f_{\theta}(\mathbf{x}_i) - y_i| - \varepsilon & |f_{\theta}(\mathbf{x}_i) - y_i| - \varepsilon > 0 \\ 0 & |f_{\theta}(\mathbf{x}_i) - y_i| - \varepsilon \leq 0 \end{cases}$$

Regularizer for Regression

- L1 regularization:

$$\Omega_1(\boldsymbol{\theta}) \propto \|\boldsymbol{\theta}\|_1 = \sum_{j=1}^m |\theta_j|$$

- L2 regularization:

$$\Omega_2(\boldsymbol{\theta}) \propto \|\boldsymbol{\theta}\|_2^2 = \sum_{j=1}^m \theta_j^2$$

Special Cases

- Lasso: squared loss + L1 regularization

$$L(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_2(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \|\boldsymbol{\theta}\|_1$$

- Ridge regression: squared loss + L2 regularization

$$L(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_2(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \|\boldsymbol{\theta}\|_2^2$$

- Elastic net: squared loss, L1 + L2 regularization

$$L(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_2(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \|\boldsymbol{\theta}\|_2^2 + \lambda' \|\boldsymbol{\theta}\|_1$$

Regularized Empirical Risk Minimization

- Solve

$$\operatorname{argmin}_{\theta} \sum_{i=1}^n \ell(f_{\theta}(\mathbf{x}_i), y_i) + \lambda \Omega(\theta)$$

- Loss function $\ell(f_{\theta}(\mathbf{x}_i), y_i)$: cost of the model's output $f_{\theta}(\mathbf{x})$ when the true value is y .
 - ◆ The empirical risk is $R_n(\theta) = \sum_{i=1}^n \ell(f_{\theta}(\mathbf{x}_i), y_i)$
 - ◆ Empirical estimate of risk $R(\theta) = \int \ell(f_{\theta}(\mathbf{x}), y) dP_{\mathbf{x},y}$
- Regularizer $\Omega(\theta)$ & trade-off parameter $\lambda \geq 0$:
 - ◆ Background information about preferred solutions
 - ◆ Provides numerical stability (Tikhonov-Regularizer)
 - ◆ allows for tighter error bounds (PAC-Theory)

Regularized Empirical Risk Minimization

- Solve

$$\operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^n \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \Omega(\boldsymbol{\theta})$$

- Linear model:

$$\operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^n \ell(\mathbf{x}_i^T \boldsymbol{\theta}, y_i) + \lambda \Omega(\boldsymbol{\theta})$$

Regularized Empirical Risk Minimization

- Linear model: solve

$$\operatorname{argmin}_{\theta} \sum_{i=1}^n \ell(\mathbf{x}_i^T \boldsymbol{\theta}, y_i) + \lambda \Omega(\boldsymbol{\theta})$$

- How to find solution:

- ◆ Classification: No analytic solution but numeric solutions (gradient descent, cutting plane, interior point method)
- ◆ Regression: analytic solution for squared loss and small number of attributes.
- ◆ Regression: numeric solutions (e.g., stochastic gradient descent) for other loss functions and for large number of attributes.

Overview

- Linear regression models
- Loss functions and regularizers for regression
- Empirical risk minimization
- Special cases:
 - ◆ Lasso
 - ◆ Ridge regression
- Analytic solution for ridge regression

Empirical Risk: Squared loss

- Squared loss function:

$$\ell_2(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) = (f_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i)^2$$

- Matrix notation of empirical risk:

$$\sum_{i=1}^n \ell_2(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) = (\mathbf{X}\boldsymbol{\theta} - \mathbf{y})^T (\mathbf{X}\boldsymbol{\theta} - \mathbf{y})$$

- Why?

$$\begin{aligned} (\mathbf{X}\boldsymbol{\theta} - \mathbf{y}) &= \begin{pmatrix} x_{11} & & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & & x_{nm} \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta}_1 \\ \vdots \\ \boldsymbol{\theta}_m \end{pmatrix} - \mathbf{y} \\ &= \begin{pmatrix} \mathbf{x}_1^T \boldsymbol{\theta} - y_1 \\ \vdots \\ \mathbf{x}_n^T \boldsymbol{\theta} - y_n \end{pmatrix} \end{aligned}$$

Lasso

- Minimize
$$L(\theta) = (\mathbf{X}\theta - \mathbf{y})^T(\mathbf{X}\theta - \mathbf{y}) + \lambda\|\theta\|_1$$
- Convex optimization criterion, only one global minimum.

Ridge Regression

- Minimize
$$L(\boldsymbol{\theta}) = (\mathbf{X}\boldsymbol{\theta} - \mathbf{y})^T(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}) + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta}$$
- Convex optimization criterion, only one global minimum.
- Analytic solution:

$$\frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta}) = 0$$

Ridge Regression

- Linear ridge regression: minimize

$$\begin{aligned}L(\boldsymbol{\theta}) &= (\mathbf{X}\boldsymbol{\theta} - \mathbf{y})^T(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}) + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta} \\&= (\boldsymbol{\theta}^T\mathbf{X}^T - \mathbf{y}^T)^T(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}) + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta} \\&= \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\theta} + \mathbf{y}^T\mathbf{y} + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta} \\&= \boldsymbol{\theta}^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}\end{aligned}$$

- Derivative:

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\theta}^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y} \\= 2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\theta} - 2\mathbf{X}^T\mathbf{y}\end{aligned}$$

Ridge Regression

- Derivative:

$$\begin{aligned}\frac{\partial}{\partial \theta} \theta^T (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}) \theta - 2 \theta^T \mathbf{X}^T \mathbf{y} + \mathbf{y}^T \mathbf{y} \\ = 2(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}) \theta - 2 \mathbf{X}^T \mathbf{y}\end{aligned}$$

- Minimum: derivative zero

$$2(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}) \theta - 2 \mathbf{X}^T \mathbf{y} = \mathbf{0}$$

Ridge Regression

- Derivative:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}^T (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}) \boldsymbol{\theta} - 2 \boldsymbol{\theta}^T \mathbf{X}^T \mathbf{y} + \mathbf{y}^T \mathbf{y} \\ = 2(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}) \boldsymbol{\theta} - 2 \mathbf{X}^T \mathbf{y}\end{aligned}$$

- Minimum: derivative zero

$$\begin{aligned}2(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}) \boldsymbol{\theta} - 2 \mathbf{X}^T \mathbf{y} &= \mathbf{0} \\ \Rightarrow (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}) &= 2 \mathbf{X}^T \mathbf{y}\end{aligned}$$

$$\Rightarrow \boldsymbol{\theta} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

Ridge Regression

- Analytic solution:

$$\boldsymbol{\theta} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

- Requires the inverse of an $m \times m$ matrix.
- Gauss-Jordan elimination procedure is $O(m^3)$.
- Coppersmith-Winograd: $O(m^{2.37\dots})$.
- Only practical for relatively small number of attributes.
- Otherwise: use stochastic gradient method (see lecture on linear classification models).

Linear Regression Models



- Loss functions and regularizers for regression.
 - ◆ Squared loss, ε -insensitive loss,
 - ◆ L2, L2 regularization.
- Empirical risk minimization
 - ◆ Analytic solution for lasso and ridge regression, only practical for limited number of attributes.
 - ◆ Stochastic gradient descent method for large-scale regression problems.