Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

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

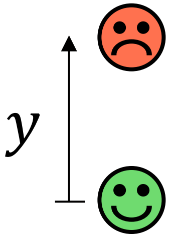
Regression

- 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*.



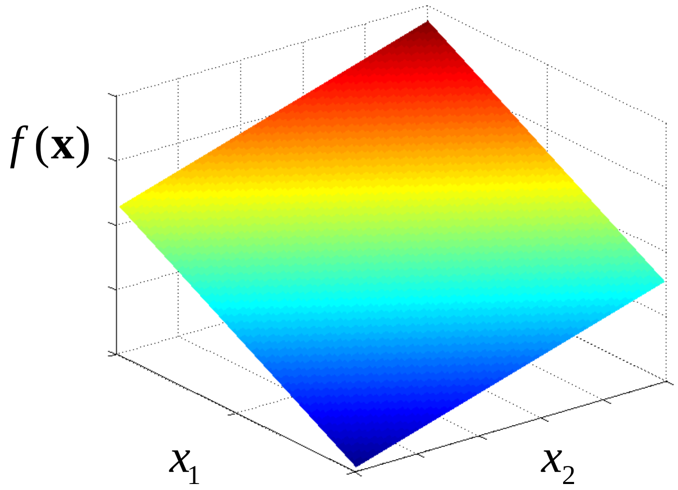
Linear Regression Models

- Regression function:

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

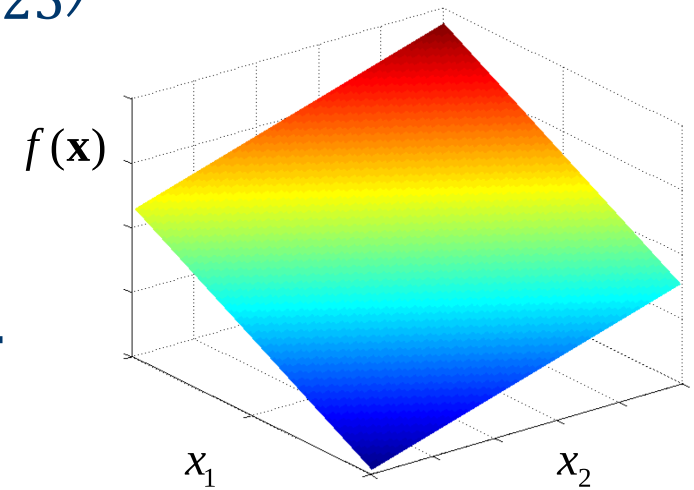
- Example:

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



Linear Regression Models

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



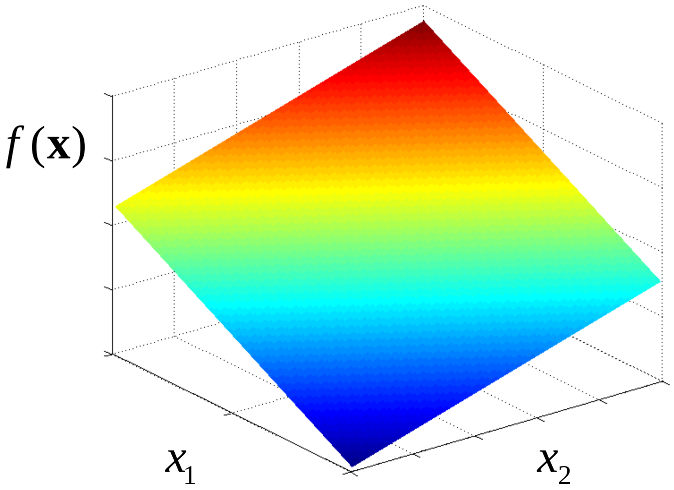
Linear Regression Models

- Regression function:

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

- Example:

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



Learning Regression Models

- Input to the Learner: Training data T_n.
  - $\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}$  
  - $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ 
- Output: a model
  -  $\mapsto$ 
    $y_{\boldsymbol{\theta}} : X \to Y$
  - For example
    $y_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}$

Linear regression model with parameter vector $\boldsymbol{\theta}$.

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

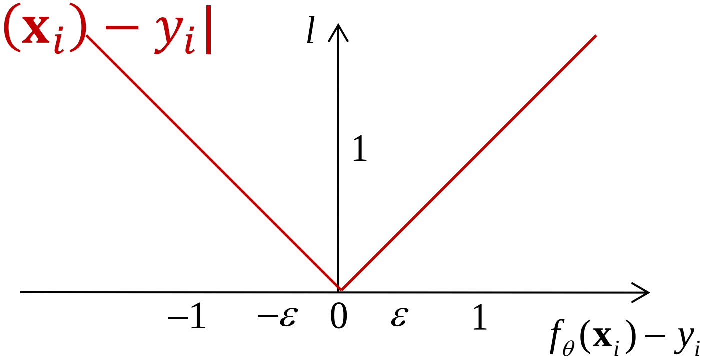
Overview

- 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:

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



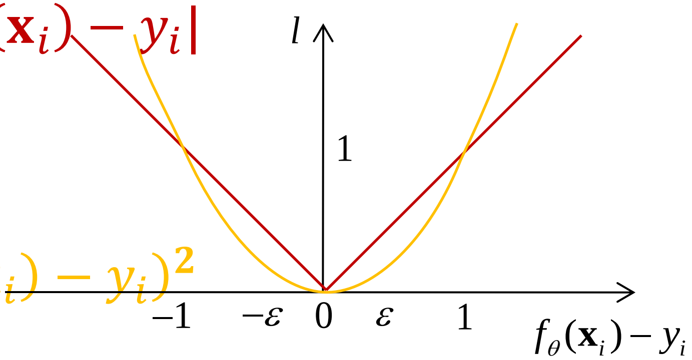
Loss Functions for Regression

- Absolute loss:

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

- Squared loss:

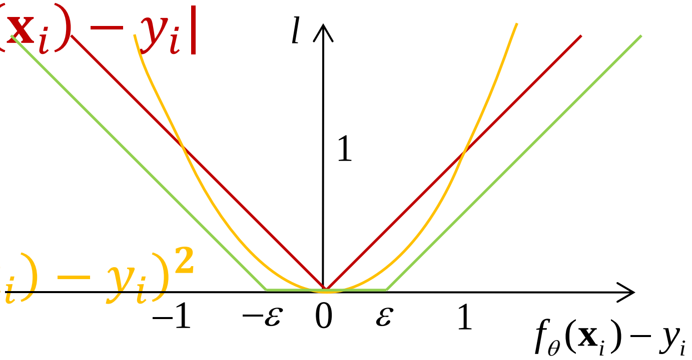
$$\ell_{\mathbf{2}}(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) = (f_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i)^{\mathbf{2}}$$



Loss Functions for Regression

- Absolute loss:

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



- Squared loss:

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

- ε-insensitive loss:

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

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

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

- Ridge regression: squared loss + L2 regularization

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

- Elastic net: squared loss, L1 + L2 regularization

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

Regularized Empirical Risk Minimization

- Solve

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

- Loss function $\ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i)$: cost of the model's output $f_{\boldsymbol{\theta}}(\mathbf{x})$ when the true value is y.
  - The empirical risk is $R_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i)$
  - Empirical estimate of risk $R(\boldsymbol{\theta}) = \int \ell(f_{\boldsymbol{\theta}}(\mathbf{x}), y) d\mathrm{P}_{\mathbf{x},y}$
- Regularizer $\Omega(\boldsymbol{\theta})$ & 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

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

- Linear model:

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

Regularized Empirical Risk Minimization

- Linear model: solve

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

- 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:

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

- Why?

$$(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}) = \begin{pmatrix} x_{11} & & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & & x_{nm} \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta}_1 \\ \vdots \\ \boldsymbol{\theta}_m \end{pmatrix} - \mathbf{y}$$

$$= \begin{pmatrix} \mathbf{x}_1{}^{\mathrm{T}}\boldsymbol{\theta} - y_1 \\ \vdots \\ \mathbf{x}_n{}^{\mathrm{T}}\boldsymbol{\theta} - y_n \end{pmatrix}$$

Lasso

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

Ridge Regression

- Minimize

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

- Convex optimization criterion, only one global minimum.
- Analytic solution:

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

Ridge Regression

- Linear ridge regression: minimize
  $$\boldsymbol{L}(\boldsymbol{\theta}) = (\mathbf{X}\boldsymbol{\theta} - \mathbf{y})^{\mathrm{T}}(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}) + \lambda\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}$$
  $$= \boldsymbol{\theta}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta} + \mathbf{y}^{\mathrm{T}}\mathbf{y} + \lambda\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}$$
  $$= \boldsymbol{\theta}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)\boldsymbol{\theta} - \mathbf{2}\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}$$
- Derivative:
  $$\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\theta}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)\boldsymbol{\theta} - \mathbf{2}\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}$$
  $$= 2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)\boldsymbol{\theta} - 2\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$$

Ridge Regression

- Derivative:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)\boldsymbol{\theta} - \mathbf{2}\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}$$
$$= 2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)\boldsymbol{\theta} - 2\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$$

- Minimum: derivative zero

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

Ridge Regression

- Derivative:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I}\right)\boldsymbol{\theta} - \mathbf{2}\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}$$
$$= 2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I}\right)\boldsymbol{\theta} - 2\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$$

- Minimum: derivative zero

$$2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I}\right)\boldsymbol{\theta} - 2\mathbf{X}^{\mathrm{T}}\boldsymbol{y} = \mathbf{0}$$
$$\Rightarrow \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I}\right)\boldsymbol{\theta} = \mathbf{X}^{\mathrm{T}}\boldsymbol{y}$$
$$\boxed{\Rightarrow \boldsymbol{\theta} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}}$$

Ridge Regression

- Analytic solution:

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

- Requires the inverse of an $m \times m$ matrix.
- Gauss-Jordan elimination procedure is $O(m^3)$.
- Coppersmith-Winograd: $O(m^{2.37\cdots})$.
- 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.