8.1 Linear Regression

Machine Learning 1: Foundations

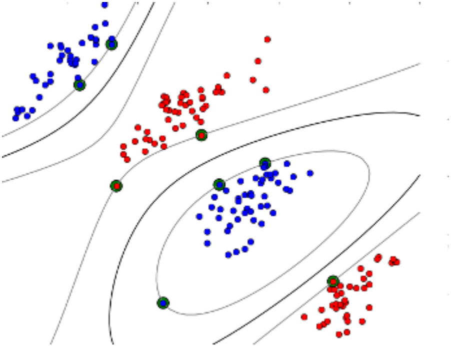
Marius Kloft *(TUK)*

Recap

In all lectures up to now, we considered **binary classification**

- meaning, the labels are binary:

$$y_1, \ldots, y_n \in \{-1, +1\}$$



Recap

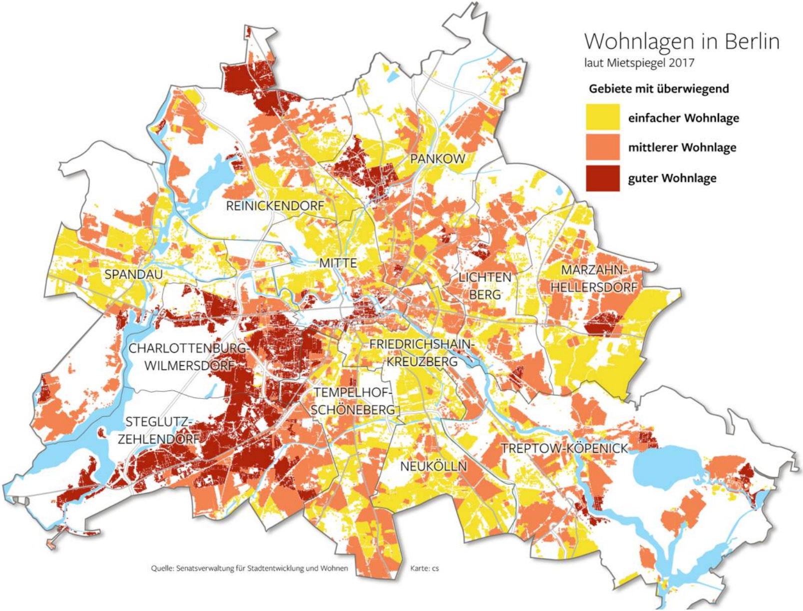
In the upcoming lectures, consider different assumptions on the labels:

- real labels (“regression”) [**today**]
- no labels (“clustering”) [next week]

Contents of this Class

Regression

Example: Rent index



Boston Housing Data Set

- Labels: median value of building (in $1000)
- Inputs: 13 features
  - AGE: proportion of owner-occupied units built prior to 1940
  - B: proportion of blacks by town
  - CRIM: per capita crime rate by town
  - DIS: weighted distances to five Boston employment centres
  - NOX: nitric oxides concentration (parts per 10 million)
  - PTRATIO: pupil-teacher ratio by town
  - RM: average number of rooms per dwelling
  - etc.

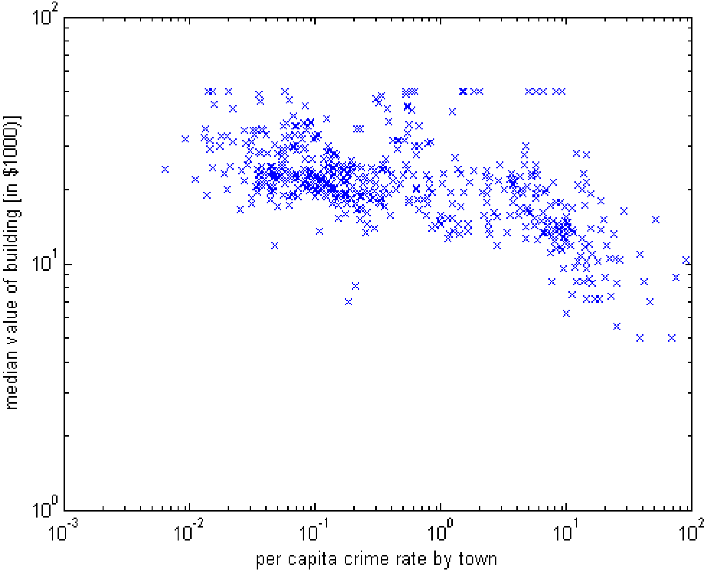
http://archive.ics.uci.edu/ml/datasets/Housing

Boston Housing Data Set

- Labels: **median value of building (in $1000)**
- Inputs: 13 features
  - AGE: proportion of owner-occupied units built prior to 1940
  - B: proportion of blacks by town
  - **CRIM: per capita crime rate by town**
  - DIS: weighted distances to five Boston employment centres
  - NOX: nitric oxides concentration (parts per 10 million)
  - PTRATIO: pupil-teacher ratio by town
  - RM: average number of rooms per dwelling
  - etc.

http://archive.ics.uci.edu/ml/datasets/Housing

The More Crime The Cheaper the House



Task

Say we own a building, how can we predict its value y from its features $\mathbf{x}$ (CRIM, AGE, etc.)?

The area of machine learning dealing with this problem is called **regression**.

Today: **Regression**

Problem setting

Given

- training inputs $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d \times n}$ and
- labels $y_1, \ldots, y_n \in \mathbb{R}$,

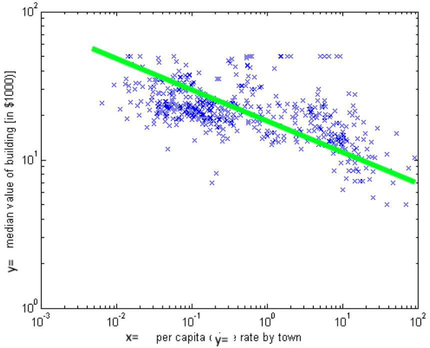
find a function $f : \mathbb{R}^d \to \mathbb{R}$ with

- $f(\mathbf{x}) \approx y$ for new data $\mathbf{x}, y$.

Key difference to *classification*:

- y is real-valued, rather than $y \in \{-1, +1\}$

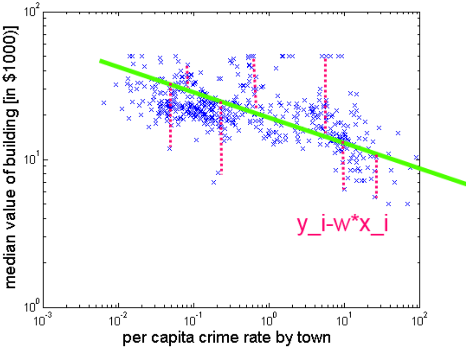
How to Predict y Given a New $\mathbf{x}$?



Linear regression: predict using a linear model $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$

But which line to take?

The Line With Minimal Distance to the Training Data



Want: $\quad y_i \approx \mathbf{w}^\top \mathbf{x}_i \quad \forall i = 1, \ldots, n$

- I.e.: $\sum_{i=1}^{n} (y_i - w^\top \mathbf{x}_i)^2 = \text{small}$
  (the hyperplane with minimal average squared distance to the training data)

The Oldest Machine Learning Method in History

Least-squares regression (Legendre, 1805)

$$\mathbf{w}_{\mathrm{LS}} := \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min} \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 = \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min} \left\| \mathbf{y} - X^\top \mathbf{w} \right\|^2$$

Definition

The function $\ell(t, y) := (t - y)^2$ is called **least-squares loss**.

Quiz: what could be a disadvantage of this method?

The Following Method is Much Better!

Idea: use a **regularizer**

Ridge regression (RR)

$$\mathbf{w}_{\mathrm{RR}} := \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\left\|\mathbf{y} - X^\top\mathbf{w}\right\|^2$$

How to compute $\mathbf{w}_{\mathrm{RR}}$?

Theorem

$$\mathbf{w}_{\mathrm{RR}} = \left(XX^\top + \frac{1}{2C}I\right)^{-1} Xy$$

Quiz: what could be a problem in practice?

- Need to compute the matrix inverse (is $O(d^3)$)

Proof

The RR problem,

$$\mathbf{w}_{\text{RR}} := \operatorname*{arg\,min}_{\mathbf{w}\in\mathbb{R}^d} \underbrace{\frac{1}{2}\|\mathbf{w}\|^2 + C\|\mathbf{y} - X^\top\mathbf{w}\|^2}_{=:\mathcal{L}(\mathbf{w})}$$

is an **unconstrained** optimization problem.

Thus optimal solution $\mathbf{w}_{\text{RR}}$ satisfies $\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}_{\text{RR}}) = 0$.

We compute (see next slide for additional details):

$$\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}_{\text{RR}}) = \mathbf{w}_{\text{RR}} - 2CX\mathbf{y} + 2CXX^\top\mathbf{w}_{\text{RR}}$$

Thus $\mathbf{w}_{\text{RR}} = \left(XX^\top + \frac{1}{2C}I\right)^{-1}Xy$ $\square$

Derivation of $\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w})$

$$
\begin{aligned}
&\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) \\
&= \nabla_{\mathbf{w}}\Big(\frac{1}{2}\,\|\mathbf{w}\|^2 + C\big\|\mathbf{y} - X^\top\mathbf{w}\big\|^2\Big) \\
&= \nabla_{\mathbf{w}}\Big(\frac{1}{2}\mathbf{w}^\top\mathbf{w} + C(\mathbf{y} - X^\top\mathbf{w})^\top(\mathbf{y} - X^\top\mathbf{w})\Big) \\
&= \nabla_{\mathbf{w}}\Big(\frac{1}{2}\mathbf{w}^\top\mathbf{w} + C(\mathbf{y}^\top - \mathbf{w}^\top X)(\mathbf{y} - X^\top\mathbf{w})\Big) \\
&= \nabla_{\mathbf{w}}\Big(\frac{1}{2}\mathbf{w}^\top\mathbf{w} + C(\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top X^\top\mathbf{w} - \mathbf{w}^\top X\mathbf{y} + \mathbf{w}^\top XX^\top\mathbf{w})\Big) \\
&= \nabla_{\mathbf{w}}\Big(\frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\mathbf{y}^\top\mathbf{y} - 2C\mathbf{w}^\top X\mathbf{y} + C\mathbf{w}^\top XX^\top\mathbf{w}\Big) \\
&= w - 2CX\mathbf{y} + 2CXX^\top\mathbf{w}
\end{aligned}
$$

What About the Bias b?

We have considered a linear model without bias:

- $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} \quad \cancel{+\mathbf{b}}$

However, we can easily incorporate a bias into any linear learning machine (regression, SVM, etc.) by the following trick:

- augment the feature space by a dimension of all ones:
  $$\forall i : \tilde{\mathbf{x}}_i := \begin{pmatrix} \mathbf{x}_i \\ 1 \end{pmatrix}, \quad \tilde{X} := (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n) = \begin{pmatrix} X \\ \mathbf{1}^\top \end{pmatrix}$$
- use $\tilde{\mathbf{w}} := \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix}$ as parameter

Example: ridge regression

$$\mathbf{w}^* := \underset{\tilde{\mathbf{w}} \in \mathbb{R}^{\boxed{d+1}}}{\arg\min} \; \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + C \left\| \mathbf{y} - \tilde{X}^\top \tilde{\mathbf{w}} \right\|^2$$

$$\underset{\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}}{\arg\min} \; \frac{1}{2} \left(\|\mathbf{w}\|^2 + b^2 \right) + C \left\| \mathbf{y} - X^\top \mathbf{w} - b\mathbf{1} \right\|^2$$

- Usually no drawback in that the bias is regularized