Machine Learning

A. G. Schwing & M. Telgarsky

University of Illinois at Urbana-Champaign, 2018

L2: Linear Regression

Last time: k-NN.

- Pros: simple (easy to implement and reason about).
- Cons: stores all data; curse of dimension.

This time: Linear regression ("ordinary least squares").

- Also simple!
- Reading: K. Murphy; Machine Learning: A Probabilistic Perspective; Chapter 7.

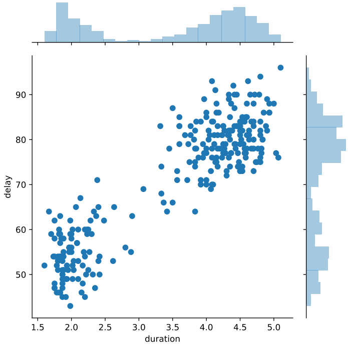
Least squares model.

Predict $y \in \mathbb{R}$ (“label”, “response”)
from $\boldsymbol{x} \in \mathbb{R}^d$ (“features”, “covariate”)
via $\boldsymbol{w}_1^\top \boldsymbol{x} + w_2$ (where $\boldsymbol{w}_1 \in \mathbb{R}^d$ and $w_2 \in \mathbb{R}$)

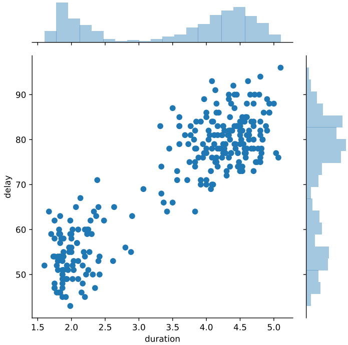
- **Learning:** choose $(\boldsymbol{w}_1, w_2)$ from data $((\mathbf{x}^{(i)}, y^{(i)}))_{i=1}^N$.
- **Prediction/inference:** obtain $\boldsymbol{x}$, output $\boldsymbol{w}_1^\top \boldsymbol{x} + w_2$.

Note. $y \in \mathbb{R}$ (“*regression*”) rather than $\{-1, +1\}$ (“*classification*”).

Example: Old faithful eruptions: duration ($\boldsymbol{x}$) vs delay (y).

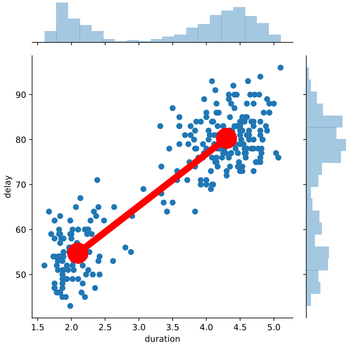


Example: Old faithful eruptions: duration ($\boldsymbol{x}$) vs delay (y).



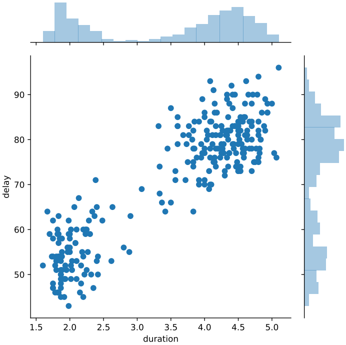
Which line $\boldsymbol{x} \mapsto \boldsymbol{w}_1^\top \boldsymbol{x} + w_2$?

Example: Old faithful eruptions: duration ($\boldsymbol{x}$) vs delay (y).



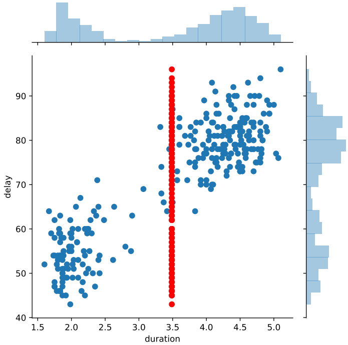
Which line $\boldsymbol{x} \mapsto \boldsymbol{w}_1^\top \boldsymbol{x} + w_2$?

Example: Old faithful eruptions: duration ($\boldsymbol{x}$) vs delay (y).



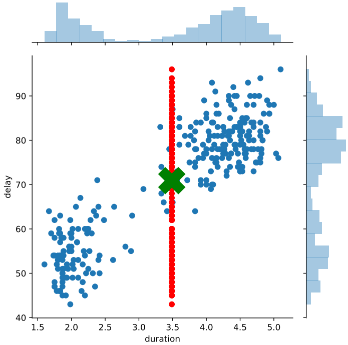
Which line $\boldsymbol{x} \mapsto \boldsymbol{w}_1^\top \boldsymbol{x} + w_2$?

Example: Old faithful eruptions: duration ($\boldsymbol{x}$) vs delay (y).



Which line $\boldsymbol{x} \mapsto \boldsymbol{w}_1^\top \boldsymbol{x} + w_2$? If all $\mathbf{x}^{(i)}$ coincide:

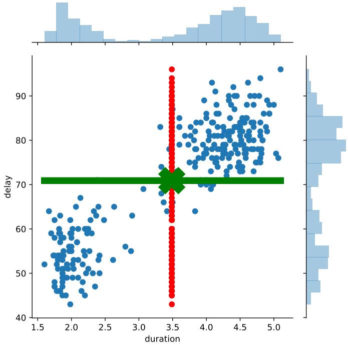
Example: Old faithful eruptions: duration ($\boldsymbol{x}$) vs delay (y).



Which line $\boldsymbol{x} \mapsto \boldsymbol{w}_1^\top \boldsymbol{x} + w_2$? If all $\mathbf{x}^{(i)}$ coincide:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} y^{(i)}$$

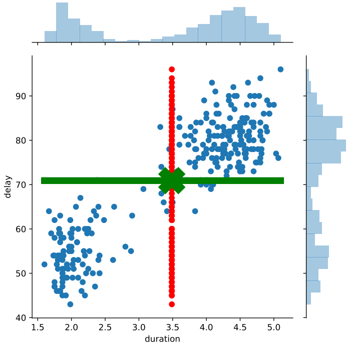
Example: Old faithful eruptions: duration ($\boldsymbol{x}$) vs delay (y).



Which line $\boldsymbol{x} \mapsto \boldsymbol{w}_1^\top \boldsymbol{x} + w_2$? If all $\mathbf{x}^{(i)}$ coincide:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} y^{(i)}$$

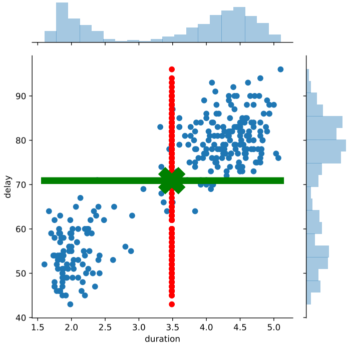
Example: Old faithful eruptions: duration ($\boldsymbol{x}$) vs delay (y).



Which line $\boldsymbol{x} \mapsto \boldsymbol{w}_1^\top \boldsymbol{x} + w_2$? If all $\mathbf{x}^{(i)}$ coincide:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} y^{(i)} = \arg\min_{y\in\mathbb{R}} \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}\left(y - y^{(i)}\right)^2 .$$

Example: Old faithful eruptions: duration ($\boldsymbol{x}$) vs delay (y).

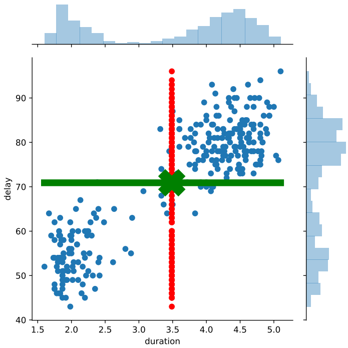


Which line $\boldsymbol{x} \mapsto \boldsymbol{w}_1^\top \boldsymbol{x} + w_2$? If all $\mathbf{x}^{(i)}$ coincide:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} y^{(i)} = \arg\min_{y \in \mathbb{R}} \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}\left(y - y^{(i)}\right)^2.$$

Remark. Mean has issues…we'll revisit this…

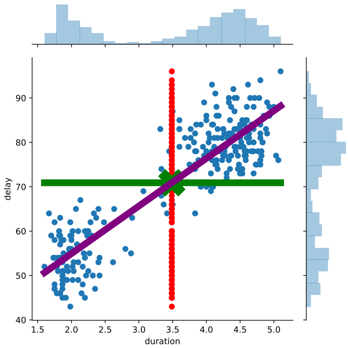
Example: Old faithful eruptions: duration ($\boldsymbol{x}$) vs delay (y).



Ignoring $\boldsymbol{x}$:

$$\arg\min_{w_2 \in \mathbb{R}} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \left(w_2 - y^{(i)} \right)^2 .$$

Example: Old faithful eruptions: duration ($\boldsymbol{x}$) vs delay (y).



Ignoring $\boldsymbol{x}$:

$$\arg\min_{w_2 \in \mathbb{R}} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \left(w_2 - y^{(i)} \right)^2 .$$

General form:

$$\arg\min_{\boldsymbol{w}_1 \in \mathbb{R}^d, w_2 \in \mathbb{R}} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \left(\boldsymbol{w}_1^\top \mathbf{x}^{(i)} + w_2 - y^{(i)} \right)^2 .$$

Example: Old faithful eruptions: duration ($\boldsymbol{x}$) vs delay (y).

Ignoring $\boldsymbol{x}$:

$$\arg\min_{w_2 \in \mathbb{R}} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \left(w_2 - y^{(i)} \right)^2 .$$

General form:

$$\arg\min_{\boldsymbol{w}_1 \in \mathbb{R}^d, w_2 \in \mathbb{R}} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \left(\boldsymbol{w}_1^\top \mathbf{x}^{(i)} + w_2 - y^{(i)} \right)^2 .$$

Example: Old faithful eruptions: duration ($\boldsymbol{x}$) vs delay (y).

Ignoring $\boldsymbol{x}$:

$$\arg\min_{w_2 \in \mathbb{R}} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \left(w_2 - y^{(i)} \right)^2 .$$

General form:

$$\arg\min_{\boldsymbol{w}_1 \in \mathbb{R}^d, w_2 \in \mathbb{R}} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \left(\boldsymbol{w}_1^\top \mathbf{x}^{(i)} + w_2 - y^{(i)} \right)^2 .$$

Simplification:

$$\mathbf{y} := \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{bmatrix}, \qquad \mathbf{X} := \begin{bmatrix} \longleftarrow \mathbf{x}^{(1)} \longrightarrow & 1 \\ \vdots & \vdots \\ \longleftarrow \mathbf{x}^{(N)} \longrightarrow & 1 \end{bmatrix}, \qquad \mathbf{w} := \begin{bmatrix} \uparrow \\ \boldsymbol{w}_1 \\ \downarrow \\ w_2 \end{bmatrix}.$$

Example: Old faithful eruptions: duration ($\boldsymbol{x}$) vs delay (y).

Ignoring $\boldsymbol{x}$:
$$\arg\min_{w_2 \in \mathbb{R}} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \left(w_2 - y^{(i)} \right)^2 .$$

General form:
$$\arg\min_{\boldsymbol{w}_1 \in \mathbb{R}^d, w_2 \in \mathbb{R}} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \left(\boldsymbol{w}_1^\top \mathbf{x}^{(i)} + w_2 - y^{(i)} \right)^2 .$$

Simplification:
$$\arg\min_{\mathbf{w} \in \mathbb{R}^{d+1}} \frac{1}{2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 \qquad \text{where}$$

$$\mathbf{y} := \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{bmatrix}, \qquad \mathbf{X} := \begin{bmatrix} \longleftarrow \mathbf{x}^{(1)} \longrightarrow & 1 \\ \vdots & \vdots \\ \longleftarrow \mathbf{x}^{(N)} \longrightarrow & 1 \end{bmatrix}, \qquad \mathbf{w} := \begin{bmatrix} \uparrow \\ \boldsymbol{w}_1 \\ \downarrow \\ w_2 \end{bmatrix}.$$

Goal: Solve

$$\arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 \qquad \text{where } \mathbf{w} \in \mathbb{R}^{d+1}, \mathbf{X} \in \mathbb{R}^{N\times(d+1)}, \boldsymbol{y} \in \mathbb{R}^{N}.$$

Goal: Solve

$$\arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \frac{1}{2}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 \qquad \text{where } \mathbf{w}\in\mathbb{R}^{d+1}, \mathbf{X}\in\mathbb{R}^{N\times(d+1)}, \boldsymbol{y}\in\mathbb{R}^{N}.$$

Ordinary least squares (OLS) estimator:
choose $\hat{\boldsymbol{w}} := (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$.

Derivation: setting derivative to 0, optimal $\hat{\boldsymbol{w}}$ satisfies

$$\boldsymbol{X}^\top(\boldsymbol{X}\hat{\boldsymbol{w}}-\boldsymbol{y}) = 0 \qquad \text{and thus} \qquad \boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{w}} = \boldsymbol{X}^\top\boldsymbol{y}.$$

When it exists, we can write $\hat{\boldsymbol{w}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$.

Non-existence of OLS solution $\hat{\boldsymbol{w}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.

Solution #1: "Ridge regression": solve

$$\arg\min_{\boldsymbol{w} \in \mathbb{R}^{d+1}} \frac{1}{2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \frac{\lambda}{2} \|\boldsymbol{w}\|^2,$$

giving $\tilde{\boldsymbol{w}} := (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.

Note. In this course's homeworks and tests, you may assume $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$ exists unless otherwise specificed.

Non-existence of OLS solution $\hat{\boldsymbol{w}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.

Solution #1: "Ridge regression": solve

$$\arg\min_{\boldsymbol{w} \in \mathbb{R}^{d+1}} \frac{1}{2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \frac{\lambda}{2} \|\boldsymbol{w}\|^2,$$

giving $\tilde{\boldsymbol{w}} := (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.

Remark. "$+\frac{\lambda}{2}\|\boldsymbol{w}\|^2$" is *regularization*;
it affects computation and statistics.

Note. In this course's homeworks and tests, you may assume $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$ exists unless otherwise specificed.

Non-existence of OLS solution $\hat{\boldsymbol{w}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.

Solution #1: "Ridge regression": solve

$$\arg\min_{\boldsymbol{w} \in \mathbb{R}^{d+1}} \frac{1}{2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \frac{\lambda}{2} \|\boldsymbol{w}\|^2,$$

giving $\tilde{\boldsymbol{w}} := (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.

Note. In this course's homeworks and tests, you may assume $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$ exists unless otherwise specificed.

Non-existence of OLS solution $\hat{\boldsymbol{w}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.

Solution #1: "Ridge regression": solve

$$\arg\min_{\boldsymbol{w} \in \mathbb{R}^{d+1}} \frac{1}{2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \frac{\lambda}{2} \|\boldsymbol{w}\|^2,$$

giving $\tilde{\boldsymbol{w}} := (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.

Solution #2: use the *pseudoinverse*:
replace $(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ with $(\boldsymbol{X}^\top \boldsymbol{X})^\dagger \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{X}^\dagger \boldsymbol{y}$.

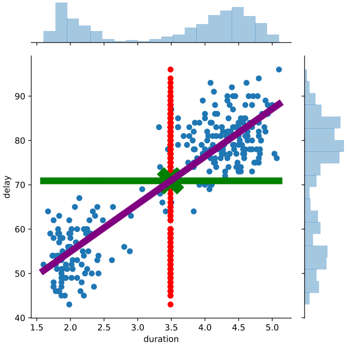
Remark. This still satisfies the "derivative condition"

$$(\boldsymbol{X}^\top \boldsymbol{X}) \hat{\boldsymbol{w}} = \boldsymbol{X}^\top \boldsymbol{y}$$

and therefore is optimal!

Note. In this course's homeworks and tests, you may assume $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$ exists unless otherwise specificed.

Summary so far



Least squares problem

$\arg\min_{\boldsymbol{w}\in\mathbb{R}^{d+1}} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$.

OLS solution

$(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$.

Question:

still “why this line” ?

Three justifications/interpretations.

- Geometric interpretation.
- Probabilistic model.
- Loss minimization.

Geometric interpretation.

Focus on **columns** of $\boldsymbol{X}$:

$$\boldsymbol{X} = \begin{bmatrix} \longleftarrow & \mathbf{x}^{(1)} & \longrightarrow & 1 \\ & \vdots & & \vdots \\ \longleftarrow & \mathbf{x}^{(N)} & \longrightarrow & 1 \end{bmatrix} = \begin{bmatrix} \uparrow & & \uparrow \\ \boldsymbol{z}_1 & \cdots & \boldsymbol{z}_{d+1} \\ \downarrow & & \downarrow \end{bmatrix}.$$

Then *residual* $\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y}$ is **orthogonal** to span $\left(\left\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{d+1}\right\}\right)$.

$(\ldots$ since $\boldsymbol{X}^\top(\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y}) = 0.)$

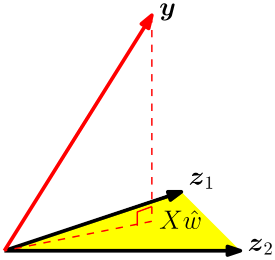
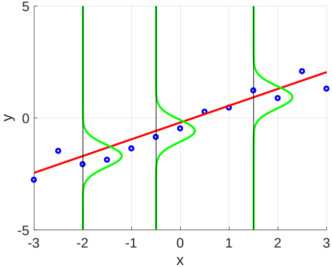


figure credit: daniel hsu

Probabilistic model.

Suppose "linear model with Gaussian errors":
label y at point $\boldsymbol{x}$ has distribution Gaussian$(\bar{\boldsymbol{w}}^\top \boldsymbol{x}, \sigma^2)$:

$$p(y^{(i)}|\boldsymbol{x}^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y^{(i)} - \bar{\boldsymbol{w}}^\top \boldsymbol{x}^{(i)})^2\right)$$



Probabilistic model.

Suppose "linear model with Gaussian errors":
label y at point $\boldsymbol{x}$ has distribution Gaussian$(\bar{\boldsymbol{w}}^\top \boldsymbol{x}, \sigma^2)$:

$$p(y^{(i)}|\boldsymbol{x}^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y^{(i)} - \bar{\boldsymbol{w}}^\top \boldsymbol{x}^{(i)})^2\right)$$

Probabilistic model.

Suppose "linear model with Gaussian errors":
label y at point $\boldsymbol{x}$ has distribution Gaussian$(\bar{\boldsymbol{w}}^\top \boldsymbol{x}, \sigma^2)$:

$$p(y^{(i)}|\boldsymbol{x}^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y^{(i)} - \bar{\boldsymbol{w}}^\top \boldsymbol{x}^{(i)})^2\right)$$

To solve, *maximize likelihood*:

$$\begin{aligned}\arg\max_{\boldsymbol{w}\in\mathbb{R}^{d+1}} \prod_{i=1}^{N} p(y^{(i)}|\mathbf{x}^{(i)}) &= (\ldots \text{hwk1} \ldots) \\ &= \arg\min_{\boldsymbol{w}\in\mathbb{R}^{d+1}} \frac{1}{2\sigma^2} \sum_{i=1}^{N} \frac{1}{2}\left(\boldsymbol{w}^\top \mathbf{x}^{(i)} - y^{(i)}\right)^2.\end{aligned}$$

Loss minimization

The form $\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2$ was convenient (OLS solution $\hat{\boldsymbol{w}} = \boldsymbol{X}^\dagger \boldsymbol{y}$).

Loss minimization

The form $\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2$ was convenient (OLS solution $\hat{\boldsymbol{w}} = \boldsymbol{X}^\dagger \boldsymbol{y}$).

Again write

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}(\boldsymbol{w}^\top \mathbf{x}^{(i)} - y^{(i)})^2 = \frac{1}{N}\sum_{i=1}^{N}\ell_{\text{ls}}(y^{(i)}, \boldsymbol{w}^\top \mathbf{x}^{(i)})$$

where now $\ell_{\text{ls}}(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$ is the *least squares loss*.

Loss minimization

The form $\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2$ was convenient (OLS solution $\hat{\boldsymbol{w}} = \boldsymbol{X}^\dagger \boldsymbol{y}$).

Again write

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}(\boldsymbol{w}^\top \mathbf{x}^{(i)} - y^{(i)})^2 = \frac{1}{N}\sum_{i=1}^{N}\ell_{\text{ls}}(y^{(i)}, \boldsymbol{w}^\top \mathbf{x}^{(i)})$$

where now $\ell_{\text{ls}}(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$ is the *least squares loss*.

The general form

$$\arg\min_f \frac{1}{N}\sum_{i=1}^{N}\ell(y^{(i)}, f(\mathbf{x}^{(i)}))$$

is the standard ML idea **Empirical Risk Minimization (ERM)**.

Three justifications/interpretations.

- Geometric interpretation.
- Probabilistic model.
- Loss minimization.

Three other questions.

- Classification vs regression.
- How to implement $\boldsymbol{X}^\dagger \boldsymbol{y}$?
- Nonlinear least squares.

Classification vs Regression.

Given $\boldsymbol{x}$, then $\hat{\boldsymbol{w}}^\top \boldsymbol{x} \in \mathbb{R}$ (“regression”)
Alternatively, $\operatorname{sgn}(\hat{\boldsymbol{w}}^\top \boldsymbol{x}) \in \{-1, +1\}$ (“classification”).

Classification vs Regression.

Given $\boldsymbol{x}$, then $\hat{\boldsymbol{w}}^\top \boldsymbol{x} \in \mathbb{R}$ ("regression")
Alternatively, $\operatorname{sgn}(\hat{\boldsymbol{w}}^\top \boldsymbol{x}) \in \{-1, +1\}$ ("classification").

Recall least squares loss $\ell_{\text{ls}}(y, \hat{y}) = (y - \hat{y})^2/2$; if $y \in \{-1, +1\}$,

$$(y - \hat{y})^2/2 = (y^2)(1 - y\hat{y})^2/2 = (1 - y\hat{y})^2/2.$$

Seems weird if our goal is to minimize $\mathbf{1}[y \neq \hat{y}] \approx \mathbf{1}[y\hat{y} \geq 0]$?

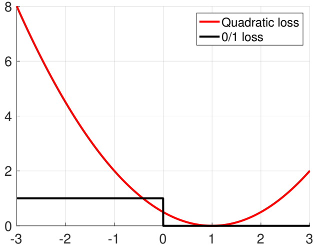
Classification vs Regression.

Given $\boldsymbol{x}$, then $\hat{\boldsymbol{w}}^\top \boldsymbol{x} \in \mathbb{R}$ ("regression")
Alternatively, $\operatorname{sgn}(\hat{\boldsymbol{w}}^\top \boldsymbol{x}) \in \{-1, +1\}$ ("classification").

Recall least squares loss $\ell_{\mathrm{ls}}(y, \hat{y}) = (y - \hat{y})^2/2$; if $y \in \{-1, +1\}$,

$$(y - \hat{y})^2/2 = (y^2)(1 - y\hat{y})^2/2 = (1 - y\hat{y})^2/2.$$

Seems weird if our goal is to minimize $\mathbf{1}[y \neq \hat{y}] \approx \mathbf{1}[y\hat{y} \geq 0]$?



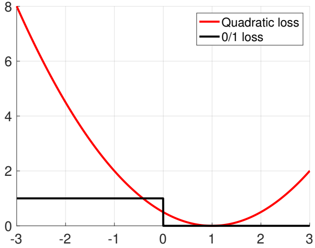
Classification vs Regression.

Given $\boldsymbol{x}$, then $\hat{\boldsymbol{w}}^\top \boldsymbol{x} \in \mathbb{R}$ ("regression")
Alternatively, $\operatorname{sgn}(\hat{\boldsymbol{w}}^\top \boldsymbol{x}) \in \{-1, +1\}$ ("classification").

Recall least squares loss $\ell_{\text{ls}}(y, \hat{y}) = (y - \hat{y})^2/2$; if $y \in \{-1, +1\}$,

$$(y - \hat{y})^2/2 = (y^2)(1 - y\hat{y})^2/2 = (1 - y\hat{y})^2/2.$$

Seems weird if our goal is to minimize $\mathbf{1}[y \neq \hat{y}] \approx \mathbf{1}[y\hat{y} \geq 0]$?



Note. Even in easy cases, linear classification is **NP**-hard!

How to solve.

Question: are $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$ and $\boldsymbol{X}^\dagger \boldsymbol{y}$ in "closed form"?

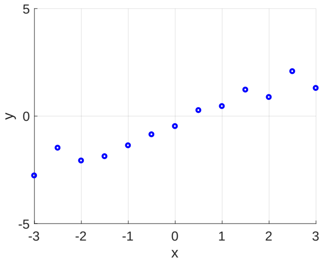
How to solve.

Question: are $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$ and $\boldsymbol{X}^\dagger \boldsymbol{y}$ in “closed form”?

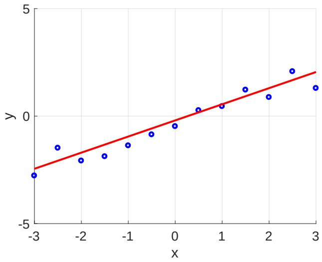
(Libraries will use iterative solvers!)

Since $\boldsymbol{w} \mapsto \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2$ is convex,
there are many “efficient” *iterative descent methods*.

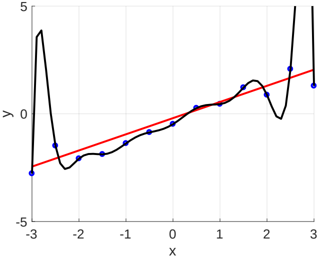
Nonlinear regression. Sometimes linear is not enough. . .



Nonlinear regression. Sometimes linear is not enough. . .

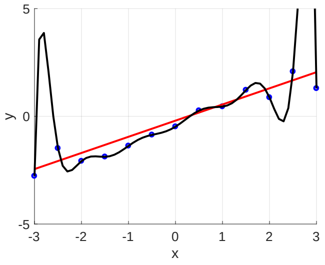


Nonlinear regression. Sometimes linear is not enough. . .



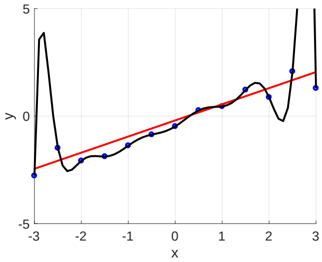
Polynomial fit?

Nonlinear regression. Sometimes linear is not enough…



Polynomial fit? (…no thanks…)

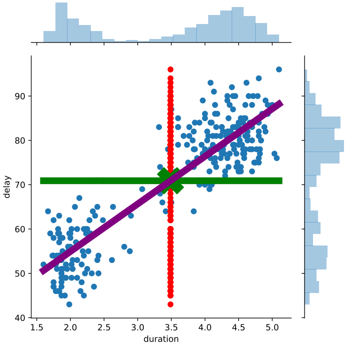
Nonlinear regression. Sometimes linear is not enough. . .



Polynomial fit? (. . . no thanks. . .)

How to solve: replace $\mathbf{x}^{(i)}$ with *features* $\widetilde{\mathbf{x}}^{(i)} = \phi(\mathbf{x}^{(i)})$.

Summary.



Least squares problem

$\arg\min_{\boldsymbol{w}\in\mathbb{R}^{d+1}} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2.$

OLS solution

$(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}.$

Justification.

- Geometric.
- Probabilistic model.
- ERM.

Concepts.

- Regularization.
- ERM and loss functions.
- Maximum likelihood.