Linear regression

CS 446 / ECE 449

2022-01-17 04:49:48 -0600 (`d748537`)

Plan for today

- Linear regression setup revisited.
- Normal equations, SVD, and pseudoinverse.
- Example (if time).

“pytorch meta-algorithm”

1. Clean/augment data. → after this step, input is vector $x \in \mathbb{R}^d$.
2. Pick model/architecture.
3. Pick a loss function measuring model fit to data.
4. Run a gradient descent variant to fit model to data.
5. Tweak 1-4 until training error is small.
6. Tweak 1-5, possibly reducing model complexity, until testing error is small.

Is that all of ML?
No, but these days it's much of it!

Linear regression — basic setup

1. Start from training data $((\boldsymbol{x}_i, y_i))_{i=1}^n$, with $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$.
2. Model is a linear predictor: pick $\boldsymbol{w} \in \mathbb{R}^d$ with

$$\boldsymbol{x}_i \mapsto \boldsymbol{w}^\top \boldsymbol{x}_i =: \hat{y}_i \approx y_i.$$

3. Loss function ℓ is squared loss ℓ_{sq} (standard regression loss):

$$\ell_{\text{sq}}(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i) = \boxed{\frac{1}{2}}(\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2.$$

makes it easy to differentiate.

We will minimize the empirical risk (average loss over training examples):

$$\widehat{\mathcal{R}}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^n \ell_{\text{sq}}(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i) = \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 \qquad \text{where } \boldsymbol{X} := \begin{bmatrix} \leftarrow & \boldsymbol{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \boldsymbol{x}_n^\top & \rightarrow \end{bmatrix}.$$

$$\frac{\partial \|Xw - y\|^2}{\partial X} = 2X^\top(Xw - y)$$

4. Basic method: gradient descent. Set $\boldsymbol{w}_0 = 0$, and thereafter

$$\boldsymbol{w}_{i+1} := \boldsymbol{w}_i - \eta \nabla \widehat{\mathcal{R}}(\boldsymbol{w}_i) = \boldsymbol{w}_i - \frac{\eta}{n}\boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{w}_i - \boldsymbol{y}),$$

where η is a learning rate (step size).

2. Model is a linear predictor: pick $\boldsymbol{w} \in \mathbb{R}^d$ with

$$\boldsymbol{x}_i \mapsto \boldsymbol{w}^\top \boldsymbol{x}_i \approx y_i.$$

- Our model/architecture/function class is $\{\boldsymbol{x} \mapsto \boldsymbol{w}^\top \boldsymbol{x} : \boldsymbol{w} \in \mathbb{R}^d\}$. For each $\boldsymbol{w} \in \mathbb{R}^d$, we have another predictor.
- This is a simple model; we'll build off of it to get more powerful ones!
- This model is insufficient for complicated tasks, but often does well, and forms a good baseline.

3. Loss function is squared loss (standard regression loss):

$$\ell(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i) = \frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2.$$

We will minimize the empirical risk:

$$\widehat{\mathcal{R}}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i) = \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2.$$

- Regression towards the mean: if $\boldsymbol{x}_i = 1 \in \mathbb{R}^1$ for all i, then

$$\arg\min_{\boldsymbol{w} \in \mathbb{R}^1} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

  Seems a reasonable notion of loss/error.
- There are many choices for ℓ. Next lecture we'll use logistic loss ℓ_{logistic}

$$\ell_{\text{logistic}}(\hat{y}, y) = \ln(1 + \exp(-\hat{y}y)).$$

  This and squared loss are the most common.

4. Basic method: gradient descent. Set $\boldsymbol{w}_0 = 0$, and thereafter

$$\boldsymbol{w}_{i+1} := \boldsymbol{w}_i - \eta\nabla\widehat{\mathcal{R}}(\boldsymbol{w}_i) = \boldsymbol{w}_i - \frac{\eta}{n}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{w}_i - \boldsymbol{y}),$$

where η is a learning rate (step size).

- In a few lectures, we'll see that this globally minimizes $\widehat{\mathcal{R}}$.
- We'll spend most of this lecture on other solutions via SVD.

Normal equations and SVD.

We want to find $\hat{\boldsymbol{w}}$ so that

$$2n\widehat{\mathcal{R}}(\hat{\boldsymbol{w}}) = \|\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y}\|^2 = \min_{\boldsymbol{w}\in\mathbb{R}^d} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 .$$

Idea from calculus: set gradient to zero and solve:

$$0 = \nabla_{\boldsymbol{w}} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 = 2\boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) ,$$

meaning we want $\hat{\boldsymbol{w}}$ so that

$$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{y}.$$

These are called the normal equations.

The normal equations are the system of linear equalities

$$\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^{\top}\boldsymbol{y}.$$

线性

Proposition. $\hat{\boldsymbol{w}}$ satisfies $\underbrace{\widehat{\mathcal{R}}(\hat{\boldsymbol{w}}) = \min_{\boldsymbol{w}} \widehat{\mathcal{R}}(\boldsymbol{w})}$ iff $\hat{\boldsymbol{w}}$ satisfies the normal equations.

Optimal training solution => global minimum.

Proof (one direction). Consider $\boldsymbol{w}$ with $\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^{\top}\boldsymbol{y}$, and any $\boldsymbol{w}'$; then

$$\begin{aligned}\|\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{y}\|^2 &= \|\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{X}\boldsymbol{w} + \boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 \\ &= \|\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{X}\boldsymbol{w}\|^2 + 2(\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{X}\boldsymbol{w})^{\top}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) + \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2.\end{aligned}$$

Since

$$(\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{X}\boldsymbol{w})^{\top}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) = (\boldsymbol{w}' - \boldsymbol{w})^{\top}(\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^{\top}\boldsymbol{y}) = 0,$$

then

$$\|\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{y}\|^2 = \|\boldsymbol{X}\boldsymbol{w}' - \boldsymbol{X}\boldsymbol{w}\|^2 + \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 \geq \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2.$$

$\square$

Later we'll get a general version by convexity, but it's nice that we can check this directly so easily!

The normal equations are the system of linear equalities

$$\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{y}.$$

Proposition. $\hat{\boldsymbol{w}}$ satisfies $\widehat{\mathcal{R}}(\hat{\boldsymbol{w}}) = \min_{\boldsymbol{w}} \widehat{\mathcal{R}}(\boldsymbol{w})$ iff $\hat{\boldsymbol{w}}$ satisfies the normal equations.

How do we solve for $\hat{\boldsymbol{w}}$?

- If $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible, we can use $(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.
- In general, we will use the SVD.

The SVD (Singular Value Decomposition).

Right. $d \times 1$

Let $\boldsymbol{M} \in \mathbb{R}^{n \times d}$ be given. $((s_i, \boldsymbol{u}_i, \boldsymbol{v}_i))_{i=1}^r$ is an SVD of $\boldsymbol{M}$ if:

Left $n \times 1$

- $\boldsymbol{M}$ has rank r;
- $s_1 \geq s_2 \cdots \geq s_r > 0$;
- $(\boldsymbol{u}_i)_{i=1}^r$ are orthonormal (orthogonal and unit length), and span the column space of $\boldsymbol{M}$;
- $(\boldsymbol{v}_i)_{i=1}^r$ are orthonormal, and span the row space of $\boldsymbol{M}$.
- $\boldsymbol{M} = \sum_i s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$. "decomposition"

- The SVD always exists, and is real-valued. (When do real eigendecompositions not exist?)
- The ordered tuple $(s_1, \ldots, s_r)$ is unique, but the SVD is in general not unique (why not?).
- For $k < r$, the low rank approximation $\sum_{i=1}^k s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top \approx \boldsymbol{M}$ has many applications (wait for the PCA lecture).

Pseudoinverse.

Given SVD $\boldsymbol{M} = \sum_i s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, the pseudoinverse is

$$\boldsymbol{M}^+ := \sum_{i=1}^{r} \frac{1}{s_i} \boldsymbol{v}_i \boldsymbol{u}_i^\top.$$

- The SVD may fail to be unique, but $\boldsymbol{M}^+$ is unique.
- $\boldsymbol{M}\boldsymbol{M}^+ = \sum_{i=1}^r \boldsymbol{u}_i \boldsymbol{u}_i^\top$ and $\boldsymbol{M}^+\boldsymbol{M} = \sum_{i=1}^r \boldsymbol{v}_i \boldsymbol{v}_i^\top$; in general, neither is an identity matrix. (Consider the case $\boldsymbol{M} = \boldsymbol{e}_1 \boldsymbol{e}_1^\top$.)

$MM^+ = \begin{bmatrix} I_{r\times r} & 0 \\ 0 & 0 \end{bmatrix}_{n\times n}$ $\quad M^+M = \begin{bmatrix} I_{r\times r} & 0 \\ 0 & 0 \end{bmatrix}_{d\times d}$

- On the other hand,

$$\boldsymbol{M}\boldsymbol{M}^+\boldsymbol{M} = \left(\sum_{i=1}^{r} s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top\right) \left(\sum_{j=1}^{r} \frac{1}{s_j} \boldsymbol{v}_j \boldsymbol{u}_j^\top\right) \left(\sum_{k=1}^{r} s_k \boldsymbol{u}_k \boldsymbol{v}_k^\top\right) = \boldsymbol{M},$$

$$\boldsymbol{M}^+\boldsymbol{M}\boldsymbol{M}^+ = \left(\sum_{i=1}^{r} \frac{1}{s_i} \boldsymbol{v}_i \boldsymbol{u}_i^\top\right) \left(\sum_{j=1}^{r} s_j \boldsymbol{u}_j \boldsymbol{v}_j^\top\right) \left(\sum_{k=1}^{r} \frac{1}{s_k} \boldsymbol{v}_k \boldsymbol{u}_k^\top\right) = \boldsymbol{M}^+.$$

- If $\boldsymbol{M}^{-1}$ exists, then $\boldsymbol{M}^+ = \boldsymbol{M}^{-1}$.
- If $\boldsymbol{M} = 0$, then $r = 0$ and $\boldsymbol{M}^+ = 0$.

OLS (Ordinary Least Squares) solution via SVD.

Given a least squares problem $\widehat{\mathcal{R}}(\boldsymbol{w}) = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2/(2n)$, the OLS solution

$$\hat{\boldsymbol{w}}_{\text{ols}} = \boldsymbol{X}^+\boldsymbol{y}$$

satisfies the normal equations (whereby $\widehat{\mathcal{R}}(\hat{\boldsymbol{w}}_{\text{ols}}) = \min_{\boldsymbol{w}} \widehat{\mathcal{R}}(\boldsymbol{w})$).

Easy to check: writing $\boldsymbol{X} = \sum_{i=1}^r s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$,

$$\begin{aligned}
\boldsymbol{X}^\top \boldsymbol{X} \hat{\boldsymbol{w}}_{\text{ols}} &= \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{X}^+ \boldsymbol{y} \\
&= \left(\sum_{i=1}^r s_i \boldsymbol{v}_i \boldsymbol{u}_i^\top\right)\left(\sum_{j=1}^r s_j \boldsymbol{u}_j \boldsymbol{v}_j^\top\right)\left(\sum_{k=1}^r \frac{1}{s_k} \boldsymbol{v}_k \boldsymbol{u}_k^\top\right)\boldsymbol{y} \\
&= \boldsymbol{X}^\top \boldsymbol{y}.
\end{aligned}$$

SVD $\boldsymbol{M} = \sum_i s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ and orthonormal bases.

We can extend $(\boldsymbol{u}_i)_{i=1}^r$ and $(\boldsymbol{v}_i)_{i=1}^r$ to full orthonormal bases for $\mathbb{R}^n$ and $\mathbb{R}^d$ respectively: write $\boldsymbol{M} \in \mathbb{R}^{n \times d}$ as

$$\left[\begin{array}{ccc|ccc} \uparrow & & \uparrow & \uparrow & & \uparrow \\ \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_r & \boldsymbol{u}_{r+1} & \cdots & \boldsymbol{u}_n \\ \downarrow & & \downarrow & \downarrow & & \downarrow \end{array}\right] \cdot \left[\begin{array}{ccc|c} s_1 & & 0 & \\ & \ddots & & 0 \\ 0 & & s_r & \\ \hline & 0 & & 0 \end{array}\right] \cdot \left[\begin{array}{ccc|ccc} \uparrow & & \uparrow & \uparrow & & \uparrow \\ \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_r & \boldsymbol{v}_{r+1} & \cdots & \boldsymbol{v}_d \\ \downarrow & & \downarrow & \downarrow & & \downarrow \end{array}\right]^\top.$$

The old parts span the column and row spaces of $\boldsymbol{M}$;
the new vectors span the left and right nullspaces.
Some call this a "full" SVD.

SVD and relationship to eigenvalues.

Note

$$\boldsymbol{M}\boldsymbol{M}^{\mathsf{T}} = \sum_{i=1}^{r} s_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}} \sum_{j=1}^{r} s_j \boldsymbol{v}_j \boldsymbol{u}_j^{\mathsf{T}} = \sum_{i=1}^{r} s_i^2 \boldsymbol{u}_i \boldsymbol{u}_i^{\mathsf{T}},$$

thus left singular vectors $(\boldsymbol{u})_{i=1}^{r}$ are top eigenvectors of $\boldsymbol{M}\boldsymbol{M}^{\mathsf{T}}$, with eigenvalues $s_1^2 \geq \cdots \geq s_r^2$.
Similarly,

$$\boldsymbol{M}^{\mathsf{T}}\boldsymbol{M} = \sum_{i=1}^{r} s_i \boldsymbol{v}_i \boldsymbol{u}_i^{\mathsf{T}} \sum_{j=1}^{r} s_j \boldsymbol{u}_j \boldsymbol{v}_j^{\mathsf{T}} = \sum_{i=1}^{r} s_i^2 \boldsymbol{v}_i \boldsymbol{v}_i^{\mathsf{T}},$$

obtaining right singular vectors from $\boldsymbol{M}^{\mathsf{T}}\boldsymbol{M}$.

Summary on least squares solutions

We want to approximately solve the empirical risk minimization problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \widehat{\mathcal{R}}(\boldsymbol{w}) = \min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|^2 .$$

Three approaches:

1. Gradient descent: $\boldsymbol{w}_0 := 0$, thereafter $\boldsymbol{w}_{i+1} := \boldsymbol{w} - \eta\nabla\widehat{\mathcal{R}}(\boldsymbol{w}_i)$.
2. Pick any $\hat{\boldsymbol{w}}$ satsifying the normal equations

$$\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^\top\boldsymbol{y}.$$

3. Use the ordinary least squares (OLS) solution $\hat{\boldsymbol{w}}_{\text{ols}} = \boldsymbol{X}^+\boldsymbol{y}$.

GD vs. OLS.

(Side note: are these different?. . .)

GD usually quicke

Both solved by iterative method, faster? depends on.

Why GD?

Why include GD, since pseudoinverse seems sufficient?

- GD is easy to implement, pseudoinverse more painful.
- Pseudoinverse after all implemented as an iterative solver.
- GD generalizes to other cases of squared loss (e.g., deep network training with squared loss).