Least squares

- So far we have looked at binary prediction problems with zero-one loss

$$\ell(h,(x,y)) = \mathbb{1}[h(x) \neq y]$$

classifier — labeled example

- Finding classifiers with optimal zero-one loss is often a hard optimization problem (e.g. NP-hard for halfspaces)

- There are ML tasks where the label y is not binary.

- We consider the setting

where the set of labels

$$\mathcal{Y} = \mathbb{R} \quad \text{and} \quad \ell(h, (x, y)) = \underbrace{(h(x) - y)^2}_{\text{square loss}}$$

- $\mathcal{Y}$: space of labels
- h: predictor, $h: \mathcal{X} \to \mathbb{R}$
- y: real-valued label

- For a labeled sample

$$S = ((x_1, y_1), \ldots, (x_m, y_m))$$

where $x_1, x_2, \ldots, x_m \in \mathcal{X}$ and

$$y_1, y_2, \ldots, y_m \in \mathbb{R}$$

we define <u>empirical loss</u>

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, (x_i, y_i))$$

- For a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ we define generalization error

we define $L_D(h)$ (risk, population loss) as

$$L_D(h) = \mathbb{E}\left[\ell(h, (x, y))\right]$$

where (x, y) is a random sample from D.

- In general, we can use any loss function

$$\ell : Y^X \times X \times Y \to \mathbb{R}^d$$

- Zero-one loss and square loss are just two examples

Linear least squares
(Ordinary least squares)

Suppose $Y = \mathbb{R}$, $X = \mathbb{R}^d$,

$$H_d = \{ h_w : w \in \mathbb{R}^d \} \subseteq Y^X$$

where $h_w : \mathbb{R}^d \to \mathbb{R}$,

$$h_w(x) = w^T x$$

and as before

$$\ell(h, (x, y)) = (h(x) - y)^2$$

ERM for linear least squares

* We are given a labeled sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ where $x_1, x_2, \ldots, x_m \in \mathbb{R}^d$ and

$y_1, y_2, \dots, y_m \in \mathbb{R}$.

- We want find the minimizer of the empirical loss:

$$\hat{h} = \underset{h \in H_d}{\operatorname{argmin}} \; L_S(h)$$

- This is equivalent to finding $\hat{w} \in \mathbb{R}^d$ such that

$$\hat{w} = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \; L_S(h_w)$$

$$= \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \; \frac{1}{m} \sum_{i=1}^{m} \left(w^T x_i - y_i \right)^2$$

- Let $f : \mathbb{R}^d \to \mathbb{R}$

$$\frac{1}{m} \sum_{i=1}^{m} \left(\quad \right)^2$$

$$f(w) = \frac{1}{m} \sum_{i=1} (w^T x_i - y_i)$$

- Then, $\hat{w}$ is the minimizer of function f.

- Since is f is differentiable, any global (or local) minimizer is a stationary point. That is, we can find $\hat{w}$ by solving

$$\nabla f(w) = 0$$

where ∇f is the gradient of f:

$$\nabla f(w) = \underbrace{\left(\frac{\partial f(w)}{\partial w_1}, \frac{\partial f(w)}{\partial w_2}, \ldots, \frac{\partial f(w)}{\partial w_d} \right)}$$

vector of partial derivatives

- The gradient is

$$\nabla f(w) = \frac{1}{m} \sum_{i=1}^{m} 2\left(w^T x_i - y_i\right) x_i$$

- The equation $\nabla f(w) = 0$ is equivalent to

$$\sum_{i=1}^{m} \left(w^T x_i - y_i\right) x_i = 0$$

- Equivalently

$$\sum_{i=1}^{m} \left(w^T x_i\right) x_i = \sum_{i=1}^{m} y_i x_i$$

$$\Updownarrow$$

$$\left(\sum_{i=1}^{m} x_i x_i^T\right) w = \sum_{i=1}^{m} y_i x_i$$

$x_i x_i^T$ is
a $d \times d$ matrix
(outer product of x_i
with itself)

- Let $A = \sum_{i=1}^{m} x_i x_i^T$ and $b = \sum_{i=1}^{m} y_i x_i$

- The last equation is equivalent to

$$A w = b$$

- Assuming A is invertible

$$w = A^{-1} b$$

- It turns out that $Aw = b$

always has a solution, even if A is not invertible. The reason is that A and b are related.

- If A is not invertible there are infinitely many solutions.

- The solution of $Aw = b$ that minimizes $\|w\|_2$ is unique and can be computed using psendo-inverse

$$w = A^{+} b$$

↑

psendo-inverse

- Since A is symmetric, it has eigen decomposition

$$A = V^T D V$$

where D is $d \times d$ diagonal matrix and V is $d \times d$ ortogonal matrix, i.e.

$$V^T = V^{-1}$$

- If $D = \begin{pmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_d \end{pmatrix}$ then the psendo-inverse of D is

$$D^+ = \begin{pmatrix} \beta_1 & & 0 \\ & \ddots & \\ 0 & & \beta_d \end{pmatrix}$$ where

$$\beta_i = \begin{cases} 1/\alpha_i & \text{if } \alpha_i \neq 0 \\ 0 & \text{if } \alpha_i = 0 \end{cases}$$

- If M is an invertible matrix then

$$M^{-1} = M^{+}$$

- Pseudo-inverse of A is

$$A^{+} = \left(V^{T} D V\right)^{+}$$

$$= V^{+} D^{+} \left(V^{T}\right)^{+}$$

$$= V^{T} D^{+} V$$

- This the same rule as for standard inverse.

$$V^+ = V^{-1} = V^T \quad \checkmark$$

$$(V^T)^+ = (V^T)^{-1} = V$$

Polynomial regression

- Let $Y = \mathbb{R}$, $X = \mathbb{R}$ and

$$H_d = \left\{ p : \begin{array}{l} p(x) \text{ is a polynomial,} \\ \deg(p) \le d \end{array} \right\}$$

$$\ell(h_1, (x,y)) = (h(x) - y)^2$$

- The problem can be reduced to linear least squares.

- Suppose

$$p(x) = w_0 + w_1 x + \cdots + w_d x^d$$

$$= w^T \psi(x)$$

where $w = (w_0, w_1, \ldots, w_d) \in \mathbb{R}^{d+1}$

and $\psi : \mathbb{R}^d \to \mathbb{R}^{d+1}$

$$\psi(x) = (1, x, x^2, \ldots, x^d)$$

- ψ is called a feature mapping

(More specifically, ψ is called polynomial feature mapping.)

- Suppose $S = ((x_1, y_1) \ldots (x_m, y_m))$

is a labeled sample where

$$x_1, x_2, \ldots, x_m \in \mathbb{R}$$

and $y_1, y_2, \ldots, y_m \in \mathbb{R}$

- The ERM problem is

$$\hat{p} = \underset{p : \deg(p) \le d}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} \left(p(x_i) - y_i \right)^2$$

- Equivalently,

$$\hat{w} = \underset{w \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} \left(w^T \psi(x_i) - y_i \right)^2$$

- We can find $\hat{w}$ using linear least squares