

Lecture 12

Linear regression

Consider $p(y|\vec{x}, \vec{\theta}) = \mathcal{N}(y | \underbrace{\vec{w} \cdot \vec{x}}_{\substack{\text{D-dim} \\ \text{vector}}} + w_0, \sigma^2)$,

where $\vec{\theta} = (\vec{w}, w_0, \sigma^2)$.

Write $\vec{x} = (1, x_1, \dots, x_D)$, absorb w_0 into $\vec{w}$: $\vec{w} \rightarrow (w_0, \vec{w})$.

Features: $p(y|\vec{x}, \vec{\theta}) = \mathcal{N}(y | \vec{w} \cdot \vec{f}(\vec{x}), \sigma^2)$

Multi-D version:

$$p(\underbrace{\vec{y}}_{\substack{\text{M-dim} \\ \text{vector}}} | \vec{x}, W) = \prod_{j=1}^M \mathcal{N}(y_j | \underbrace{\vec{w}_j \cdot \vec{x}}_{\substack{\text{D+1} \\ \text{dim}}} , \sigma_j^2)$$

↑
factorizes into M 1D problems

with correlations,

$$p(\vec{y} | \vec{x}, W) = \mathcal{N}(\vec{y} | \underbrace{W \cdot \vec{x}}_{\substack{\text{M} \times (\text{D}+1)}}, \underbrace{\sum_{\text{M} \times \text{M}}}_{\substack{\uparrow \\ \text{covariance} \\ \text{matrix}}})$$

—○—

Next, consider the 1D case:

$$\begin{aligned} -\mathcal{L}(\vec{w}, \sigma^2) &= \frac{1}{2\sigma^2} \sum_{n=1}^N (y_n - \vec{w} \cdot \vec{x}_n)^2 + \\ &+ \frac{N}{2} \log(2\pi\sigma^2). \end{aligned}$$

Focus first on

[i.e., infer $\vec{w}$ first,
 σ^2 second]

$$RSS(\vec{w}) = \frac{1}{2} \sum_{n=1}^N (y_n - \vec{w} \cdot \vec{x}_n)^2 \quad \ominus$$

↑
residual sum
of squares

$$\ominus \frac{1}{2} (\vec{y} - X\vec{w}) \cdot (\vec{y} - X\vec{w}).$$

↑
dim = N

$$\underbrace{X}_{N, D+1} \underbrace{\vec{w}}_{D+1} = N\text{-dim vector.}$$

data matrix, each row is
an input pattern (augmented with 1)

Note that

$$\underbrace{\frac{\partial}{\partial w_i} RSS(\vec{w})}_{D+1\text{-dim vector}} = - \sum_{n=1}^N \underbrace{(y_n - \vec{w} \cdot \vec{x}_n)}_{\substack{\vec{y} - X\vec{w} \\ N\text{-dim}}} \underbrace{x_{n,i}}_{X_{D+1,N}^T}$$

Then

$$+ \frac{\partial}{\partial \vec{w}} RSS(\vec{w}) = \underbrace{X^T X \vec{w}}_{(D+1) \times (D+1)} - X^T \vec{y} = 0, \text{ s.t.}$$

$\equiv \vec{g}(\vec{w})$

$$\vec{w} = \underbrace{(X^T X)^{-1} X^T \vec{y}}_{\text{Moore-Penrose pseudoinverse}}. \quad (*)$$

moreover, $\frac{\partial^2}{\partial w_i \partial w_j} RSS(\vec{w}) = \sum_n x_{n,i} x_{n,j},$

$\hookrightarrow X^T X$

$$H(\vec{w}) = \underline{\underline{X^T X}}.$$

Now, for $\forall \vec{v} > 0$ we have

$$\vec{v}^T (X^T X) \vec{v} = (X \vec{v})^T (X \vec{v}) > 0 \text{ if}$$

X is full-rank.

So, $H(\vec{w})$ is pos.-def., and the $\vec{g}(\vec{w})=0$ solution is unique and is given by Eq. (*).

Typically, $N \gg D \Rightarrow$ the X matrix is 'tall and skinny'. This is an overdetermined system of linear equations.

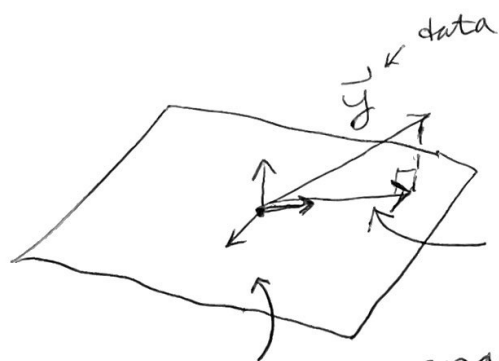
Geometrically, we seek to represent an N -dim vector $\vec{y} = \underline{y_1, y_2, \dots, y_N}$ as a linear combination of $D+1$ N -dim vectors $\vec{x}_1^c, \dots, \vec{x}_{D+1}^c \Leftarrow$ columns of X .

These vectors span $(D+1)$ -dim subspace of the N -dim space if X is full rank.

In our model, we consider $X \vec{w} =$

$$= w_1 \vec{x}_1^c + \dots + w_{D+1} \vec{x}_{D+1}^c \text{ as our prediction.}$$

We want to minimize the magnitude of the difference between $\vec{y}$ & $X \vec{w}$.



($b=0$ for this example)

optimal model $X\bar{w}^*$, orthogonal projection onto M

$M = \text{subspace spanned by } \bar{x}_1^c, \dots, \bar{x}_{D+1}^c, \text{ embedded into } \mathbb{R}^N$

Clearly, $\bar{x}_i^c \cdot (\bar{y} - X\bar{w}^*) = 0, \forall i=1, \dots, D+1$

Then $X_{D+1, N}^T (\bar{y} - X\bar{w}^*) = \vec{0}_{D+1}$, or

$$(X^T X) \bar{w}^* = X^T \bar{y},$$

$$\bar{w}^* = (X^T X)^{-1} X^T \bar{y}, \text{ same as Eq. (*) above}$$

If we denote $\hat{\bar{y}} = X\bar{w}^*$,

$$\hat{\bar{y}} = X(X^T X)^{-1} X^T \bar{y}$$

projection operator

Numerically, the main issue in using Eq. (*) is to invert $X^T X$ or using SVD or other methods. $\left[\begin{array}{l} \text{compute} \\ (X^T X)^{-1} X \end{array} \right]$

Weighted least squares

Consider $p(y|\vec{x}, \vec{\theta}) = \mathcal{N}(y|\vec{w} \cdot \vec{x}, \underbrace{\sigma^2(\vec{x})}_{\substack{\text{variance} \\ \text{depends on} \\ \text{input } \vec{x}}})$

Then $p(\vec{y}|\vec{x}, \vec{\theta}) = \mathcal{N}(\vec{y}|X\vec{w}, \Lambda^{-1})$, where
" $\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$ $\Lambda = \text{diag}[\frac{1}{\sigma^2(\vec{x}_1)}, \dots, \frac{1}{\sigma^2(\vec{x}_N)}]$

$$RSS(\vec{w}) = \underbrace{\frac{1}{2} \sum_{n=1}^N (y_n - \vec{w} \cdot \vec{x}_n)^2 \frac{1}{\sigma^2(\vec{x}_n)}}_{\text{weighted sum of squares}}.$$

If the weights are known, we obtain:

$$\frac{\partial}{\partial w_i} RSS(\vec{w}) = - \sum_n (y_n - \vec{w} \cdot \vec{x}_n) \frac{x_{n,i}}{\sigma^2(\vec{x}_n)}.$$

Then, if $\frac{\partial}{\partial \vec{w}} RSS(\vec{w}) = 0 = (X^T \Lambda X) \vec{w} - X^T \Lambda \vec{y}$,
or

$$\underline{\underline{\vec{w}^* = (X^T \Lambda X)^{-1} X^T \Lambda \vec{y}}}.$$

— Sometimes it makes sense to compute the bias and the D-dim weights separately.

Specifically, consider

$$RSS(\vec{w}) = \frac{1}{2} \sum_{n=1}^N (y_n - \overset{w_0}{b} - \underbrace{\vec{w} \cdot \vec{x}_n}_{D\text{-dim}})^2.$$

Then $\frac{\partial}{\partial b} RSS(\vec{w}) = -\sum_{n=1}^N (y_n - b - \vec{w} \cdot \vec{x}_n) = 0$, or

$$Nb = \underbrace{\sum_n y_n}_{N\langle y \rangle} - \vec{w} \cdot \underbrace{\sum_n \vec{x}_n}_{N\langle \vec{x} \rangle},$$

(**) $\left[b = \underbrace{\langle y \rangle}_{\text{ave of data}} - \underbrace{\vec{w} \cdot \langle \vec{x} \rangle}_{\substack{D\text{-dim} \\ \text{ave of each component} \\ \text{of } \vec{x} \text{ over } N \text{ datapoints}}} \right]$

Next, $RSS(\vec{w}) = \frac{1}{2} \sum_{n=1}^N (\underbrace{y_n - \langle y \rangle}_{y_n^c} - \vec{w} \cdot (\underbrace{\vec{x}_n - \langle \vec{x} \rangle}_{\vec{x}_n^c}))$.

Same as before, but with $D\text{-dim } \vec{x}$ and both $\vec{x}$ & y 'centered' by subtracting the mean.

Then $\vec{w}_D^* = (X_c^T X_c)^{-1} X_c^T \vec{y}_c$ and

b is found using Eq. (**).

Note that here

$$\vec{w}^* = \underbrace{\left[\sum_{n=1}^N (\vec{x}_n - \langle \vec{x} \rangle)(\vec{x}_n - \langle \vec{x} \rangle)^T \right]^{-1}}_{D \times D \text{ matrix}} \times \underbrace{\sum_{n=1}^N (y_n - \langle y \rangle)(\vec{x}_n - \langle \vec{x} \rangle)}_{D\text{-dim vector}}.$$

$$\mathcal{H}_0: D=1,$$

$$\begin{cases} w_1^* = \frac{\sum_n (x_n - \langle x \rangle)(y_n - \langle y \rangle)}{\sum_n (x_n - \langle x \rangle)(x_n - \langle x \rangle)} = \frac{\text{Cov}(x, y)}{\text{Var}(x)}, \\ b = \langle y \rangle - w_1^* \langle x \rangle. \end{cases}$$

Recall that

$$-\mathcal{I}(\vec{w}, \sigma^2) = \frac{\text{RSS}(\vec{w})}{\sigma^2} + \frac{N}{2} \log(2\pi\sigma^2).$$

Then

$$\frac{\partial}{\partial \sigma^2} (-\mathcal{I}(\vec{w}, \sigma^2)) = -\frac{\text{RSS}(\vec{w})}{(\sigma^2)^2} + \frac{N}{2\sigma^2} = 0, \quad \text{or}$$

$$\sigma^2 = \frac{2}{N} \text{RSS}(\vec{w}) = \frac{1}{N} \sum_{n=1}^N (y_n - \vec{w} \cdot \vec{x}_n)^2.$$

=

MSE of the residuals