Lecture 16

Linear regression

Consider $p(y|\vec{x},\vec{\theta}) = \mathcal{N}(y|\vec{w}\cdot\vec{x}+w_0,\sigma^2)$,

($\vec{x}$: D-dim vector)

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

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

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

Multi-D version:

$$p(\vec{y}|\vec{x},W) = \prod_{j=1}^{M} \mathcal{N}(y_j|\vec{w}_j\cdot\vec{x},\sigma_j^2)$$

($\vec{y}$: M-dim vector; j; $\vec{w}_j$: D+1 dim)

⇑ factorizes problems into M 1D

with correlations,

$$p(\vec{y}|\vec{x},W) = \mathcal{N}(\vec{y}|\underbrace{W}_{M\times(D+1)}\cdot\vec{x}, \underbrace{\Sigma}_{M\times M})$$

(Σ: covariance matrix)

Next, consider the 1D case:

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

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

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

↑ residual sum of squares

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

$\vec{y}$: dim = N

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

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

Note that

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

Then

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

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

Moore-Penrose pseudoinverse

Moreover, $\frac{\partial^2}{\partial w_i \partial w_j} RSS(\vec{w}) = \underbrace{\sum_n x_{n,i}\, x_{n,j}}_{\to X^T X}$,

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

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

$$\vec{v}^T (X^T X) \vec{v} = (X\vec{v})^T (X\vec{v}) > 0$$ 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. (*).

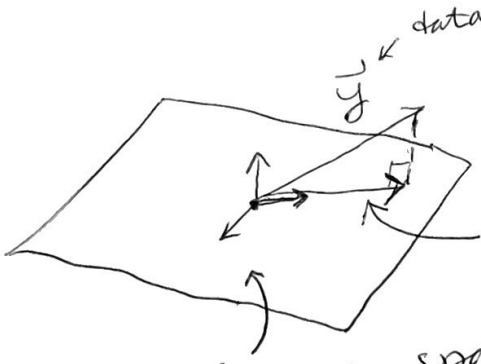
Typically, $N \gg D$ $\Rightarrow$ the X matrix is 'tall and skinny'. This is an overdetermined system of linear equations.

Geometrically, we seek to represent an N-dim vector $\vec{y} = y_1, y_2, \ldots, y_N$ as a linear combination of D+1 N-dim vectors $\vec{x}^c_1, \ldots, \vec{x}^c_{D+1} \Leftarrow$ 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 $X\vec{w} =$

$= w_1 \vec{x}^c_1 + \ldots + w_2 \vec{x}^c_{D+1}$ as our prediction.

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



($b=0$ for this example)

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

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

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

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

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

$$\hat{\vec{y}} = \underbrace{X(X^T X)^{-1} X^T}_{\text{projection operator}} \vec{y}$$

Numerically, the main issue in using Eq. (*) is to invert $X^T X$ or using SVD or other methods. [compute $(X^T X)^{-1} X$]

Weighted least squares

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

Then $p(\vec{y}|\vec{x},\vec{\theta}) = \mathcal{N}(\vec{y}|X\vec{w}, \Lambda^{-1})$, where

"$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$" $\quad \Lambda = \text{diag}\left[\frac{1}{\sigma^2(\vec{x}_1)}, \dots, \frac{1}{\sigma^2(\vec{x}_N)}\right]$.

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

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

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

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

Specifically, consider w_0

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

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

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

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

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

Same as before, but with D-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

$$\vec{w}^* = \Big[\underbrace{\sum_{n=1}^{N}(\vec{x}_n - \langle\vec{x}\rangle)(\vec{x}_n - \langle\vec{x}\rangle)^T}_{D\times D \text{ matrix}}\Big]^{-1} \times$$

$$\times \underbrace{\sum_{n=1}^{N}(y_n - \langle y\rangle)(\vec{x}_n - \langle\vec{x}\rangle)}_{D\text{-dim vector}} .$$

If $D=1$,

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

Recall that

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

Then

$$\frac{\partial}{\partial \sigma^2}\left(-\mathcal{L}(\vec{w}, \sigma^2)\right) = -\frac{RSS(\vec{w})}{(\sigma^2)^2} + \frac{N}{2\sigma^2} = 0,$$

or

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