Linear models for regression — Lecture 2

Simplest model:

$$y(\vec{x}, \vec{w}) = w_0 + \sum_{j=1}^{D} w_j x_j$$

$\{\vec{x}_1 \ldots \vec{x}_N\}$ multi-D observations $\rightarrow$ each $\vec{x} = \{x_1 \ldots x_D\}$

$\{t_1 \ldots t_N\}$ target variables

More generally,

$$y(\vec{x}, \vec{w}) = \underbrace{w_0}_{\text{bias prm}} + \sum_{j=1}^{D} w_j \underbrace{\varphi_j(\vec{x})}_{\text{basis f's}}$$

still a linear model because y is linear in $\vec{w}$

Sometimes, one defines $\varphi_0(\vec{x}) = 1$, s.t.

$$y(\vec{x}, \vec{w}) = \sum_{j=0}^{D} w_j \varphi_j(\vec{x}) = \vec{w}^T \cdot \vec{\varphi}(\vec{x})$$

$$\vec{w} = (w_0 \ldots w_D) \qquad \vec{\varphi} = (\varphi_0 \ldots \varphi_D)$$

Popular choices of basis f's:
(1D examples)

(1) Gaussian $\varphi_j(x) = e^{-\frac{(x-\mu_j)^2}{2s^2}}$

(2) Sigmoidal $\varphi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right)$, where

$$\sigma(a) = \frac{1}{1+e^{-a}}$$

As before, we assume that

$$t = y(\vec{x}, \vec{w}) + \xi$$

$\uparrow$ target $\quad$ $\uparrow$ model $\quad$ $\uparrow$ noise

$$p(\xi \mid \beta) = \mathcal{N}(0, \beta^{-1})$$

Correspondingly,

$$p(t \mid \vec{x}, \vec{w}, \beta) = \mathcal{N}(t \mid y(\vec{x}, \vec{w}), \beta^{-1})$$

Now consider

$$\vec{X} = \{\underbrace{\vec{x}_1, \dots, \vec{x}_N}_{\text{inputs}}\}$$

$$\vec{t} = \{\underbrace{t_1, \dots, t_N}_{\text{target vars}}\}$$

Under the independence assumption,

$$p(\vec{t} \mid \vec{X}, \vec{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid \vec{w}^T \cdot \vec{\varphi}(\vec{x}_n), \beta^{-1})$$

drop conditional dependence on $\vec{X}$ for brevity

Then

$$\log p(\vec{t} \mid \vec{w}, \beta) = \frac{N}{2} \log \beta - \frac{N}{2} \log(2\pi) - \beta \underbrace{\frac{1}{2} \sum_{n=1}^{N} (t_n - \vec{w}^T \vec{\varphi}(\vec{x}_n))^2}_{E_D(\vec{w})} \quad (**)$$

$$= \sum_{n=1}^{N} \log \mathcal{N}(t_n \mid \vec{w}^T \vec{\varphi}(x_n), \beta^{-1})$$

Now, use ML to determine $\vec{w}_{ML}$ & β_{ML}.
Start with the weights:

$$\frac{\partial}{\partial w_j} \log P(\vec{t} \mid \vec{w}, \beta) = \frac{\beta}{\cancel{2}} \cancel{2} \sum_{n=1}^{N} \left(t_n - \vec{w}^T \vec{\varphi}(\vec{x}_n) \right) \times \varphi_j(\vec{x}_n),$$

or

$$\underbrace{\vec{\nabla}_{\vec{w}} \log P}_{=0} = \beta \sum_{n=1}^{N} \left(t_n - \vec{w}^T \vec{\varphi}(\vec{x}_n) \right) \vec{\varphi}(\vec{x}_n)$$

$$\Rightarrow \sum_{n=1}^{N} t_n \varphi_j(\vec{x}_n) = \sum_{i=0}^{D} w_i \sum_{n=1}^{N} \varphi_i(\vec{x}_n) \varphi_j(\vec{x}_n)$$

Call $\varphi_j(\vec{x}_n) = \Phi_{nj}$ → elements of $N \times (D+1)$ <u>design</u> <u>matrix</u>

Then

$$\underbrace{\overbrace{\sum_n t_n \Phi_{nj}}^{(\Phi^T \vec{t})_j}}_{\substack{D+1 \\ \text{vector}}} = \sum_i w_i \underbrace{\sum_n \Phi_{ni} \Phi_{nj}}_{\substack{(D+1)\times(D+1) \\ \text{matrix}}} = \underbrace{\sum_i (\Phi^T \Phi)_{ji} w_i}_{\substack{D+1 \\ \text{vector}}}$$

Finally,

$$\Phi^T \vec{t} = (\Phi^T \Phi) \vec{w}, \quad \text{or}$$

$$\vec{w}_{ML} = (\Phi^T \Phi)^{-1} \Phi^T \vec{t}$$

normal eq's for the least-squares problem

$$\Phi = \begin{pmatrix} \varphi_0(\vec{x}_1) & \varphi_1(\vec{x}_1) & \dots & \varphi_{M-1}(\vec{x}_1) \\ \vdots & \vdots & & \vdots \\ \varphi_0(\vec{x}_N) & \varphi_1(\vec{x}_N) & \dots & \varphi_{M-1}(\vec{x}_N) \end{pmatrix}$$

$\tilde{\Phi} = (\Phi^T\Phi)^{-1}\Phi^T$ is called the Moore-Penrose pseudo-inverse of Φ.

Next, consider

$$E(\vec{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(t_n - w_0 - \sum_{j=1}^{D} w_j \varphi_j(\vec{x}_n)\right)^2.$$

$$\frac{\partial E}{\partial w_0} = -\sum_{n=1}^{N}\left(t_n - w_0 - \sum_{j=1}^{D} w_j \varphi_j(\vec{x}_n)\right) = 0$$

gives

$$w_0^{ML} = \frac{1}{N}\sum_{n=1}^{N} t_n - \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{D} w_j \varphi_j(\vec{x}_n) =$$

$$= \bar{t} - \sum_{j=1}^{D} w_j \bar{\varphi}_j \quad , \text{ where } \bar{\ } \text{ indicates } \frac{1}{N}\sum_{n=1}^{N}\ldots$$

Thus w_0^{ML} is the difference between the average of the target values and the weighted sum of basis f'n averages.

Finally, maximize (**) wrt β:

$$\frac{1}{\beta_{ML}} = \underbrace{\frac{1}{N}\sum_{n=1}^{N}\left(t_n - \vec{w}_{ML}^T\,\vec{\varphi}(\vec{x}_n)\right)^2}$$

residual variance of the target values around $y(\vec{x}, \vec{w})$

Geometry of least squares

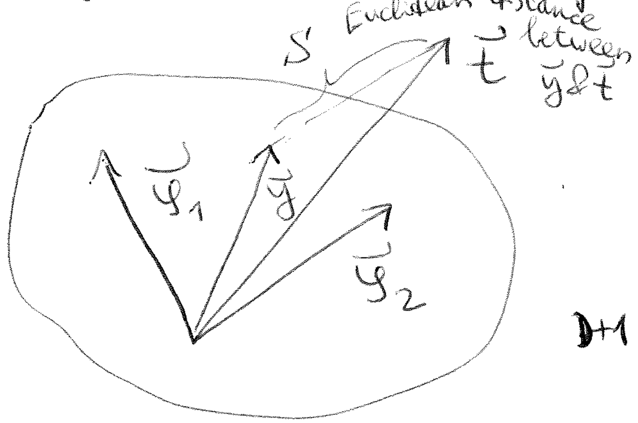
$\vec{t} = (t_1 \dots t_N)$ is an N-dim vector

Each $\vec{\varphi}_j = (\varphi_j(\vec{x}_1) \dots \varphi_j(\vec{x}_N))$ is also an N-dim vector

[$\vec{\varphi}_j$ is the j^{th} column of Φ]

If $\underbrace{D+1}_{\text{the \# of basis functions, including } \varphi_0(\vec{x}_n)} < \overbrace{N}^{\#\text{datapoints}}$, then $\vec{\varphi}_j$ vectors

will span a subspace S' of dimensionality $D+1$ in the N-dim'l space.



Define $\vec{y} = (y(\vec{x}_1, \vec{w}) \dots y(\vec{x}_N, \vec{w}))$
N-dim vector, a linear combination of $D+1$ $\vec{\varphi}_j$ vectors.

$E(\vec{w}) \sim$ ~~second~~ square of Euclidean dist. between $\vec{y}$ & $\vec{t}$, so we need to minimize it. This is accomplished by the orthogonal projection of $\vec{t}$ into subspace S'.

In practice, $\Phi^T\Phi$ may be hard to invert, when some $\vec{\varphi}_j$'s are nearly colinear $\rightarrow$ add a regularization term to prevent this.

Regularized least squares

Consider now

$$\tilde{E}(\vec{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(t_n - \vec{w}^T \vec{\varphi}(\vec{x}_n)\right)^2 + \frac{\lambda}{2}\vec{w}^T\vec{w}$$

still a quadratic f'n of w_j

$\vec{w}_{ML} = (\Phi^T\Phi)^{-1}\Phi^T\vec{t}$ becomes

$\vec{w}^* = (\lambda \mathbf{I} + \Phi^T\Phi)^{-1}\Phi^T\vec{t}$ [the inverse matrix is now regularized]

$\mathbf{I}$: unit matrix $(D+1)\times(D+1)$

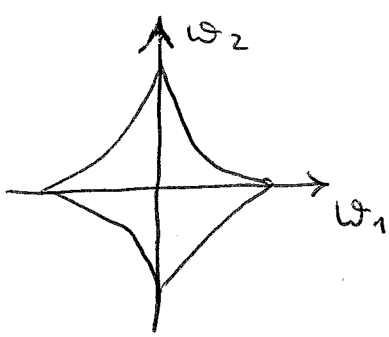
More generally,

$$\tilde{E}(\vec{w}) = \frac{1}{2}\sum_{n=1}^{N}(\ldots)^2 + \frac{\lambda}{2}\sum_{j=1}^{D+1}|w_j|^q$$

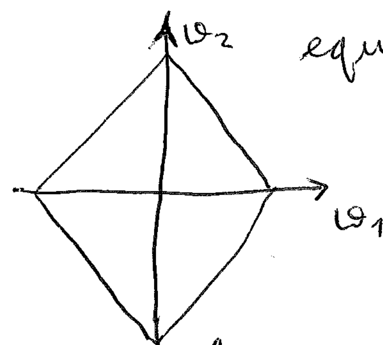
$q = 2$ ← quadratic regularizer above

$q = 1$ ← LASSO regularizer

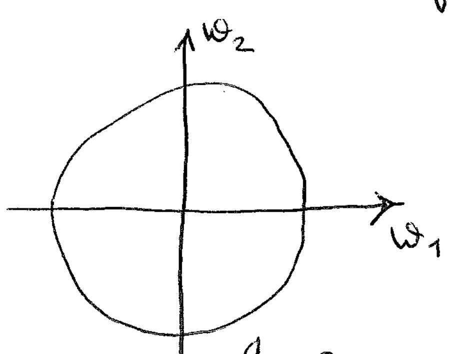
equal value contours



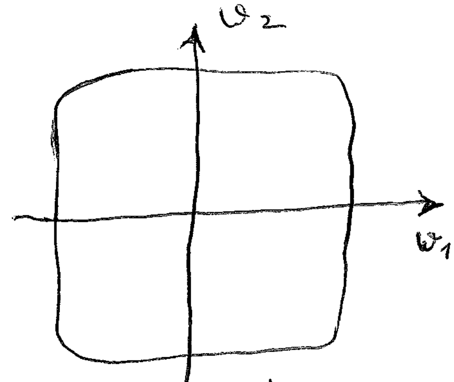
$q=0.5$



$q=1$

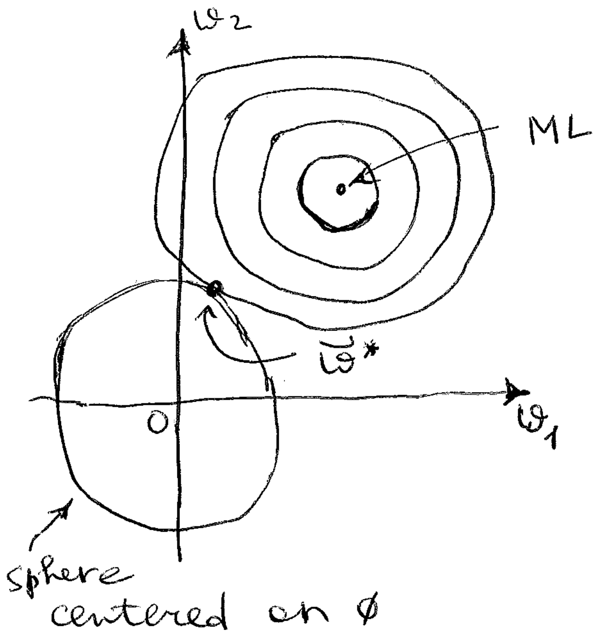


$q=2$

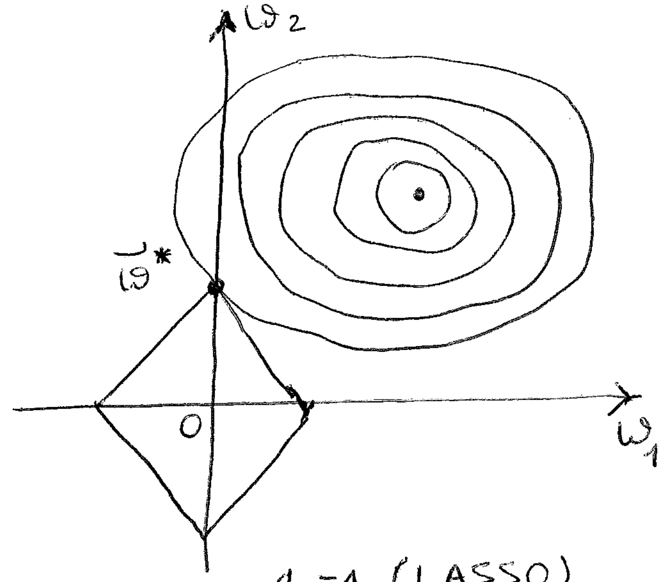


$q=4$

Now, consider



$q=2$



$q=1$ (LASSO) results in (sparse solution)

Loss functions for regression (1.5.5)

Consider $\underset{\text{input}}{\vec{x}} \to \underset{\text{target}}{t}$

$$\underset{\substack{\text{model}\\ \text{for } t}}{y(\vec{x})} \to \underset{\text{loss}}{L(t, y(\vec{x}))} \overset{\text{e.g.}}{=} \underbrace{(y(\vec{x})-t)^2}_{\text{squared loss}}$$

$$\underset{\substack{\text{expected}\\ \text{value of}\\ \text{loss}}}{E[L]} = \int d\vec{x}\, dt\, \underbrace{p(\vec{x},t)}_{\substack{\text{joint prob.}\\ \text{of } \vec{x} \,\&\, t}} \underbrace{(y(\vec{x})-t)^2}_{\text{loss f'n}}$$

$$\frac{\delta E[L]}{\delta y(x)} = 2\int dt\, (y(\vec{x})-t)\, p(\vec{x},t) = 0$$

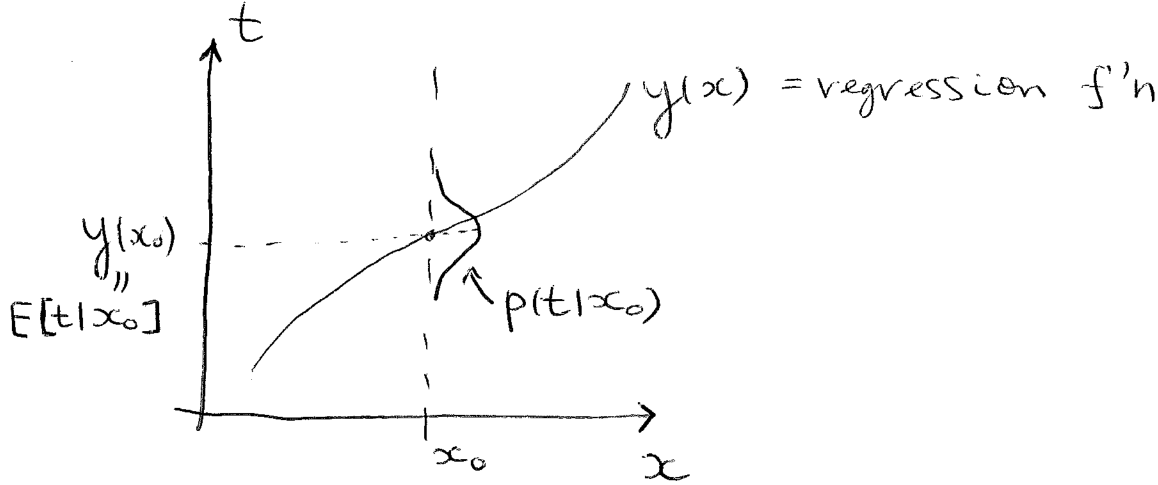
minimize $E[L]$

$$\Downarrow$$

$$y(\vec{x}) = \frac{\int dt\, p(\vec{x},t)\, t}{p(\vec{x})} = \int dt\, t\, p(t|\vec{x}) =$$

$$= E[t|\vec{x}] \qquad (+)$$

$y(\vec{x})$ is called a <u>regression</u> <u>f'n</u>



Further,

$$(y(\vec{x})-t)^2 = (y(\vec{x}) - E[t|\vec{x}] + E[t|\vec{x}] - t)^2 =$$

$$= (y(\vec{x}) - E)^2 + 2(y(\vec{x}) - E)(E - t) + (E-t)^2$$

Then
$$E[L] = \int dt\, d\vec{x}\, p(\vec{x},t)\Big[(y(\vec{x}) - \overbrace{E[t|\vec{x}]}^{\text{f'n of } \vec{x}\text{, \underline{not} } t})^2 +$$

$$+ 2(y(\vec{x}) - E[t|\vec{x}])(E[t|\vec{x}]-t) + (E[t|\vec{x}]-t)^2\Big] =$$

$$= \int d\vec{x}\, p(\vec{x})\, (y(\vec{x}) - E[t|\vec{x}])^2 +$$

$$+ 2\int d\vec{x}\, (y(\vec{x}) - E[t|\vec{x}]) \underbrace{\int dt\, p(\vec{x},t)(E[t|\vec{x}]-t)}_{E[t|\vec{x}]\,p(\vec{x}) \;-\; p(\vec{x})\,E[t|\vec{x}] \;=\; 0} +$$

2nd term vanishes

$$+ \int dt\, d\vec{x}\, p(\vec{x},t)\,(E[t|\vec{x}]-t)^2 .$$

Only the <u>1st</u> term depends on $y(\vec{x})$, and will be $=0$ (at min) iff

$$y(\vec{x}) = E[t|\vec{x}]\,,$$ consistent with (+)

The 3rd term represents noise in the target values & is indep. of the model:

$$\int d\vec{x}\, \underbrace{p(\vec{x})\, E^2[t|x]} - 2\int dt\, d\vec{x}\; t\, p(\vec{x},t)\, E[t,\vec{x}] +$$

$$+ \int dt\, t^2 p(t) =$$

$$= \int d\vec{x}\; \frac{\int dt\, t\, p(\vec{x},t) \int dt'\, t'\, p(\vec{x},t')}{p(\vec{x})} -$$

$$-2\int dt\, d\vec{x}\; t\, p(\vec{x},t)\, \frac{\int dt'\, t'\, p(\vec{x},t')}{p(\vec{x})} + \int dt\, t^2 p(t) =$$

$$= \langle t^2 \rangle - \underbrace{\int \frac{d\vec{x}}{p(\vec{x})} \int dt\, t\, p(\vec{x},t) \int dt'\, t'\, p(\vec{x},t')}_{\int d\vec{x}\, p(\vec{x}) \left(\underbrace{\frac{\int dt\, t\, p(\vec{x},t)}{\int dt\, p(\vec{x},t)}}_{= \langle t \rangle \; (\text{f'n of } \vec{x})} \right)^2} .$$

$$\int dt\, t^2 p(t) = \frac{\int dt\, d\vec{x}\; t^2 p(\vec{x},t)}{\underbrace{\int dt\, d\vec{x}\; p(\vec{x},t)}_{=1}} .$$