Today (Linear Regression ctd)

→ Statistical intuition on Ridge and LASSO
(distinction between MLE (Maximum Likelihood) and the MAP (Maximum A Posteriori estimator))

→ Comparison between Ridge and LASSO through constrained formulations (ℓ_q balls)

→ Effect of regularization (Ridge & LASSO) on bias and variance

→ Plot bias and variance decomposition for the MSE

→ Classification

→ Simple least squares binary classifier

↳ Extensions to multiple classes fitting

→ logistic regression

Statistical Motivation for Ridge and LASSO

$$\ell_{Ridge}(\beta) = \frac{1}{N}\sum_{i=1}^{N}(t^{(i)} - \beta^T \tilde{x}^{(i)})^2 + \lambda \sum_{j=1}^{D} |\beta_j|^2$$

$\tilde{x}^{(i)} \in \mathbb{R}^{D+1}$

$\tilde{x}^{(i)} = [1, x^{(i)}]$

$\underbrace{\{t^{(i)}, x^{(i)}\}_{i=1}^{N}}$ training data

Assumption 1

$$t^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)} + \varepsilon^{(i)}$$

Assumption 2 $\varepsilon^{(i)} \sim \mathcal{N}(0, \sigma)$

$\mathcal{N}(0, \sigma)$ = Gaussian distn.

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Assumption 3 $\varepsilon^{(i)}$ independent

(i.i.d independent & identically distributed)

Since the only random contribution to $t^{(i)}$ is $\varepsilon^{(i)}$ the probability of observing a particular $t^{(i)}$ can be deduced from $p(\varepsilon^{(i)})$. In particular $\varepsilon^{(i)} = t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})$

$$p(t^{(i)} \mid x^{(i)}; \beta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ - \frac{(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}))^2}{2\sigma^2} \right\} \quad (*)$$

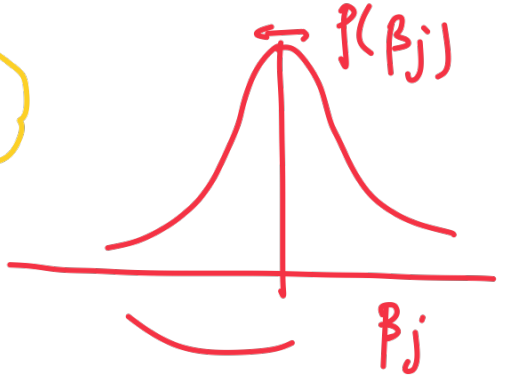
→ Now assume that we have some prior intuition on how the β_j (parameters we want to estimate) should look like.

and assume that this intuition is given to us in the form of a probability distribution $p(\beta_j)$

Further assume that β_j's are independent. (Gaussian mean 0 variance λ^2)

As an example let us take $p(\beta_j) = \frac{1}{\sqrt{2\pi}\lambda} \exp\left(-\frac{(\beta_j)^2}{2\lambda^2}\right)$ (*)

$p(\beta) \overset{\Delta}{=} \prod_{j=1} p(\beta_j)$



Question? How can we incorporate this additional prior information on β_j's in order to improve our MLE estimator?

→ Use Bayes' theorem

posterior distribut (*)

$$p(\beta \mid t^{(i)}) = \frac{p(t^{(i)} \mid \beta)\, p(\beta)}{p(t^{(i)})}$$ (*)

$p(t^{(i)})$

$$p(t^{(i)}) = \sum_{\forall \beta_k} p(t^{(i)} \mid \beta_k)\, p(\beta_k)$$ → constant across different choices of β_k

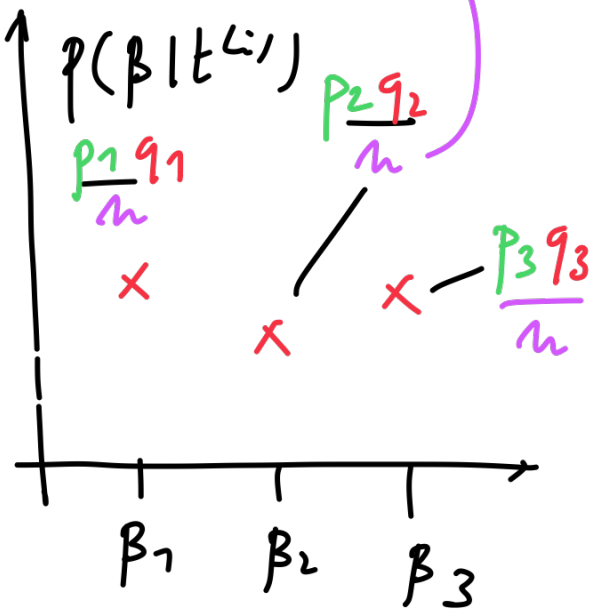
A good estimator for β is given by

Maximizing the posterior

$$\underset{\beta}{\operatorname{argmax}}\; p(\beta \mid \{t^{(i)}, x^{(i)}\}_{i=1}^{N})$$

$$= \underset{\beta}{\operatorname{argmax}}\; p(\{t^{(i)}, x^{(i)}\}_{i=1}^{N} \mid \beta)\, p(\beta)$$

$$n = p(t^{(i)} \mid \beta_1)\, p(\beta_1) + p(t^{(i)} \mid \beta_2)\, p(\beta_2) + p(t^{(i)} \mid \beta_3)\, p(\beta_3)$$



using the log to simplify the product and the exponentials.

$$\underset{\beta}{\operatorname{argmax}}\; \log\left\{ \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}))^2}{2\sigma^2} \right) \times \prod_{j=1}^{D} \frac{1}{\sqrt{2\pi}\lambda} \exp\left(-\frac{\beta_j^2}{2\lambda^2} \right) \right\}$$

$$\underset{\beta}{\operatorname{argmax}} \; p(\{t^{(i)}, x^{(i)}\}_{i=1}^{N} \mid \beta)\, p(\beta)$$

$$= \underset{\beta}{\operatorname{argmax}} \; \sum_{i=1}^{N} \left(\cancel{\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)} - \frac{1}{2\sigma^2}\left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})\right)^2 \right)$$

$$\left\{ + \sum_{j=2}^{D} \left(\cancel{\log\left(\frac{1}{\sqrt{2\pi}\lambda}\right)} - \frac{1}{2\lambda^2}\beta_j^2 \right) \right.$$

do not depend on β_j

$$= \underset{\beta}{\operatorname{argmax}} \; -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})\right)^2 - \frac{1}{2\lambda^2} \sum_{j=1}^{D} \beta_j^2 \quad *$$

Maximizing $-\ell(\beta)$ is equivalent to minimizing $\ell(\beta)$ from this

$$(*) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}) \right)^2 + \frac{\sigma^2}{\lambda^2} \sum_{j=1}^{D} \beta_j^2 = \underline{\beta_{MAP}}$$

$\longrightarrow$ Under our assumptions on $\varepsilon^{(i)}$, β_j, finding β_{Ridge} is equivalent to finding the MAP estimator for β

(Maximization of posterior distribution)

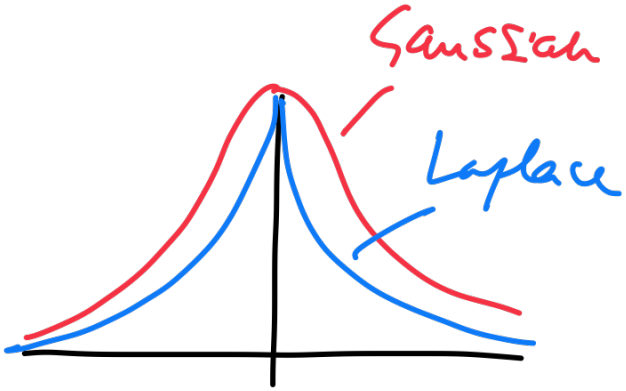
In short: MLE : β is found by maximizing the likelihood function

MAP : Some prior information on β (encoded in $p(\beta)$) is used to define a posterior distribution which is then maximized

With respect to β.

→ the Gaussian prior is good at reducing the variance as it corresponds to the assumption that β_j are most likely small but we can do better.

→ In the Gaussian, there is not a big difference between being 0 and being very small.



To further emphasize the feature selection of the model (i.e the fact that the β_j should be exactly set to 0 whenever possible

We can use Spikier distributions (e.g. Laplace)

→ Applying the previous (MAP) steps we get

$$\prod_{j=1}^{D} \frac{1}{2b} \exp\left(-\frac{|\beta_j|}{\sigma}\right)$$

$$\beta_{MAP} = \operatorname{argmax}_{\beta} \; p(\{t^{(i)}, x^{(i)}\}_{i=1}^{N} \mid \beta)\, p(\beta)$$

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}))^2}{2\sigma^2}\right)$$

$$\beta_{MAP} = \operatorname*{argmin}_{\beta} \sum_{i=1}^{N} \left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})\right)^2 + \frac{\sigma^2}{b} \sum_{j=1}^{D} |\beta_j|$$

For a Laplace prior, the MAP estimator is the LASSO loss.

minimizer of the

In short: under appropriate statistical assumptions,

β_{Ridge} = MAP with Gaussian prior

β_{LASSO} = MAP with Laplace prior

Geometric intuition

Ridge $\ell(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})\right)^2 + \lambda \underbrace{\sum_{j=1}^{D} \beta_j^2}_{(*)}$

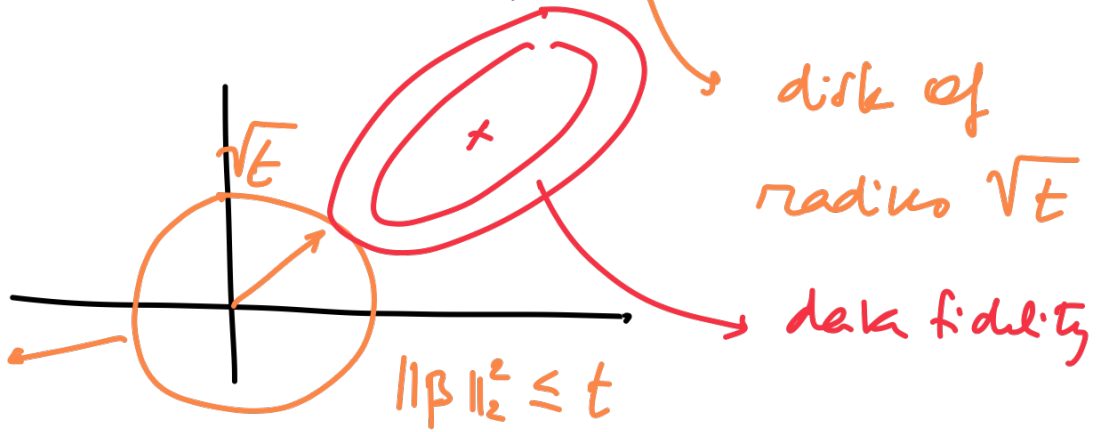
LASSO $\ell(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})\right)^2 + \lambda \sum_{j=1}^{D} |\beta_j|$

To get some additional intuition on how each formulation handles feature selection, let us consider the equivalent constrained formulations

$$\beta_{Ridge} = \min_{\beta} \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})\right)^2$$

$$\text{s.t} \quad \sum_{j=1}^{D} |\beta_j|^2 \leq t \qquad (**)$$

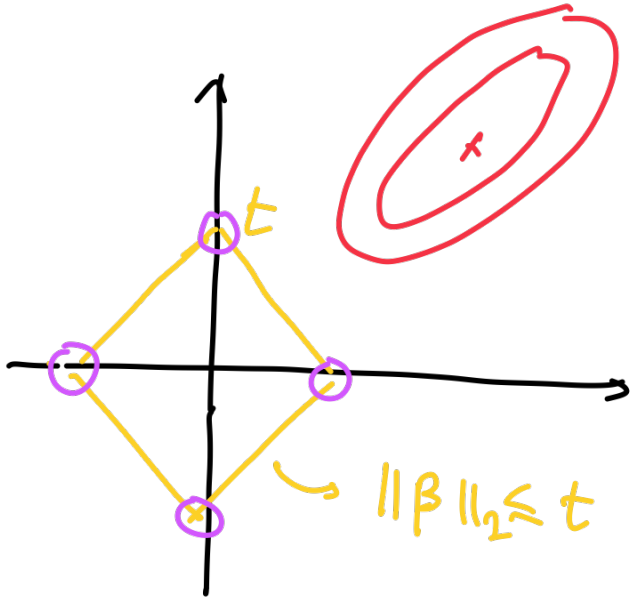
a larger value of t in $(**)$ corresponds to a smaller λ in $(*)$



Such a constrained formulation can be equivalently defined for LASSO

$$\beta_{LASSO} = \min \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}) \right)^2$$

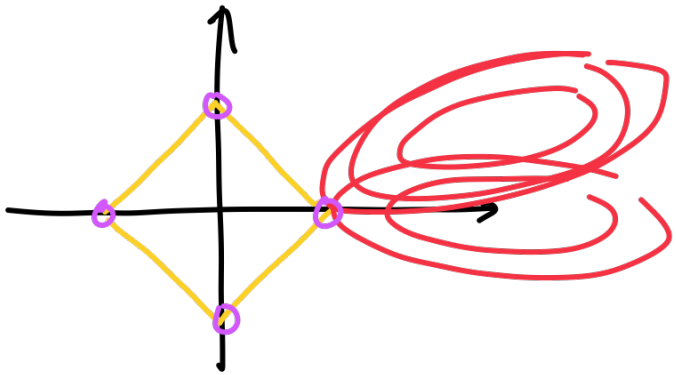
$$s.t \quad \sum_{j=1}^{D} |\beta_j| \leq t$$

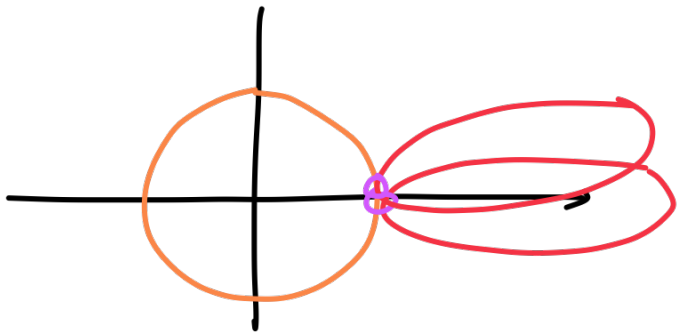


Both LASSO and Ridge can be understood as looking for the β that achieve the smallest possible values of the data fidelity term (objective corresponding to the red parabolas) within a certain ball (l_1 ball → LASSO

l_2 ball → Ridge

→ Because the l_1 ball is more spiky (such as Laplace vs Gauss) the β inside the ball achieving the minimum of the data fidelity term is more likely to be located on the intersection with the axes than it is in the l_2 ball





Both Ridge and LASSO rely on a definition of the model complexity penalty based on the l_p norm for some p

In the case of Ridge we minimize the squared l_2 norm

$$\Omega(\beta) = \sum_{j=1}^{D} \beta_j^2 = \|\beta\|_2^2$$

In the case of LASSO we minimize the l_1 norm

$$\Omega(\beta) = \sum_{j=1}^{D} |\beta_j| = \|\beta\|_1$$

Of course other penalties can be defined for other values of p

$$\Omega(p) = \sum_{j=1}^{D} |\beta_j|^p = \|\beta\|_p^p = \left(\sqrt[p]{\sum_{j=1}^{D} |\beta_j|^p}\right)^p$$