Linear regression

Alexander Sirotkin

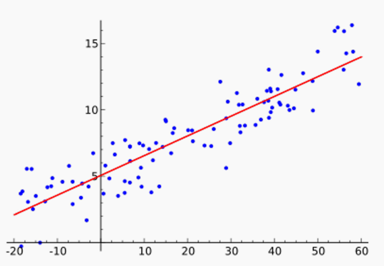
HSE University

November 14, 2023

Linear regression

- Linear model: consider a linear function

$$y(\mathrm{x}, \mathrm{w}) = w_0 + \sum_{j=1}^{p} x_j w_j = \mathrm{x}^\top \mathrm{w}, \quad \mathrm{x} = (1, x_1, \ldots, x_p).$$



- How can we find optimal parameters $\hat{\mathrm{w}}$ by training data of the form $(\mathrm{x}_i, y_i)_{i=1}^{N}$?

Linear regression

- How can we find optimal parameters $\hat{\mathrm{w}}$ by training data of the form $(\mathrm{x}_i, y_i)_{i=1}^{N}$?
- Least squares estimation: we will minimize

$$\mathrm{RSS}(\mathrm{w}) = \sum_{i=1}^{N} (y_i - \mathrm{x}_i^\top \mathrm{w})^2.$$

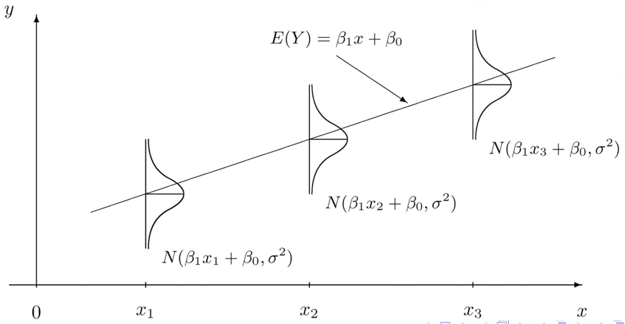
- There is even an exact solution, but that's not important right now.

Linear regression

- What is important: suppose that noise (error in the data) has a normal distribution, i.e., observed variable t is

$$t = y(\mathrm{x}, \mathrm{w}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \text{ то есть}$$

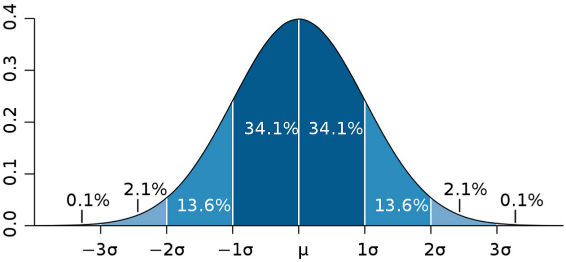
$$p(t \mid \mathrm{x}, \mathrm{w}, \sigma^2) = \mathcal{N}(t \mid y(\mathrm{x}, \mathrm{w}), \sigma^2).$$



Linear regression

- Aside – normal distribution:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$



- Why is it so important?

Linear regression

- Consider a dataset $\mathrm{X} = \{\mathrm{x}_1, \dots, \mathrm{x}_N\}$ with correct answers $\mathrm{t} = \{t_1, \dots, t_N\}$.
- We assume that the data points are independent identically distributed:

$$p(\mathrm{t} \mid \mathrm{X}, \mathrm{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \mathrm{w}^\top \phi(\mathrm{x}_n), \sigma^2\right).$$

- We take the logarithm (we omit X below for brevity):

$$\ln p(\mathrm{t} \mid \mathrm{w}, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(t_n - \mathrm{w}^\top \phi(\mathrm{x}_n)\right)^2.$$

Linear regression

- And we see that to maximize the likelihood w.r.t. w we need to minimze mean squared error!

$$\nabla_{\mathrm{w}} \ln p(\mathrm{t} \mid \mathrm{w}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^{N} \left(t_n - \mathrm{w}^\top \phi(\mathrm{x}_n) \right) \phi(\mathrm{x}_n).$$

- We can also get a posterior distribution, introducing prior distributions (also normal).
- And then the predictive distribution

$$p(y \mid \mathrm{x}, D) = \int_{\mathrm{w}} p(y \mid \mathrm{x}, \mathrm{w}) p(\mathrm{w} \mid D) \mathrm{d}\mathrm{w}$$

  ...but that's beside the point right now.
- Main conclusion: in many regression problems it makes sense to minimize the sum of squared deviations, this corresponds to normally distributed noise.

Bayesian regularization

- And now let us look at regression from the pure Bayesian perspective.
- Recall that in Bayesian inference, we
  1. find the posterior distribution on на hypothesis/params:

     $$p(\theta \mid D) \propto p(D|\theta)p(\theta)$$

     (and/or find the maximal a posteriori hypothesis $\arg\max_\theta p(\theta \mid D)$);
  2. find the predictive distribution:

     $$p(x \mid D) \propto \int_{\theta\in\Theta} p(x \mid \theta)p(D|\theta)p(\theta)\mathrm{d}\theta.$$

Bayesian regularization

- We have not yet had any priors in our study of linear regression.
- Let us introduce a prior; e.g., the normal distribution:

$$p(\mathrm{w}) = \mathcal{N}(\mathrm{w} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0).$$

- Consider a dataset $\mathrm{X} = \{\mathrm{x}_1, \ldots, \mathrm{x}_N\}$ with values $\mathrm{t} = \{t_1, \ldots, t_N\}$; we again assume that they are independent and identically distributed:

$$p(\mathrm{t} \mid \mathrm{X}, \mathrm{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \mathrm{w}^\top \phi(\mathrm{x}_n), \sigma^2\right).$$

Bayesian regularization

- Then the problem is to compute

$$p(\mathrm{w} \mid \mathrm{t}) \propto p(\mathrm{t} \mid \mathrm{X}, \mathrm{w}, \sigma^2) p(\mathrm{w})$$
$$= \mathcal{N}(\mathrm{w} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \mathrm{w}^\top \phi(\mathrm{x}_n), \sigma^2\right).$$

- Let us compute!

Bayesian regularization

- We get

$$p(\mathrm{w} \mid \mathrm{t}) = \mathcal{N}(\mathrm{w} \mid \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N),$$
$$\mu_N = \boldsymbol{\Sigma}_N \left(\boldsymbol{\Sigma}_0^{-1} \mu_0 + \frac{1}{\sigma^2} \boldsymbol{\Phi}^\top \mathrm{t} \right),$$
$$\boldsymbol{\Sigma}_N = \left(\boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma^2} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \right)^{-1}.$$

- And now the log likelihood.

Bayesian regularization

- If we take the prior distribution around zero:

$$p(\mathrm{w}) = \mathcal{N}(\mathrm{w} \mid 0, \frac{1}{\alpha}\mathrm{I}),$$

we get the log likelihood as

$$\ln p(\mathrm{w} \mid \mathrm{t}) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(t_n - \mathrm{w}^\top \phi(\mathrm{x}_n)\right)^2 - \frac{\alpha}{2}\mathrm{w}^\top \mathrm{w} + \mathrm{const},$$

i.e., precisely ridge regression!

Generalization

- A slight generalization – a more general prior distribution:

$$p(\mathrm{w} \mid \alpha) = \left[\frac{q}{2}\left(\frac{\alpha}{2}\right)^{1/q}\frac{1}{\Gamma(1/q)}\right]^M e^{-\frac{\alpha}{2}\sum_{j=1}^M |w_j|^q}.$$

Try to compute the log likelihood.

Regularization again

- We know that least squares do not always work well. Two reasons:
  1. bad predictive power – often better to regularize, trading bias for variance;
  2. hard to interpret – we often want to understand what is going on, and if we have lots of different nonzero numbers, it's hard.
- Hence, we'd like to get more nonzero components in the vector w.

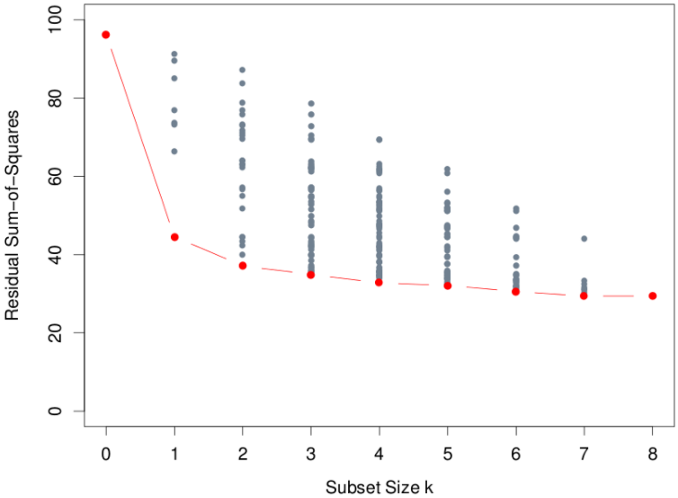
Subset selection

- What if we do it directly? Simply presume most coefficients are zero and find the nonzero ones.
- This is called *subset selection*.
- Best subset selection: choose the subset of k input variables that gives the best results

Subset selection

- Naturally, this does not work computationally: there are lots of subsets.
- Forward-stepwise selection: start from the intercept, then add one best predictor per step.
- Backward-stepwise selection: start from full regression and remove the predictor that influences the error the least.

Subset selection



Lasso

- Let us now consider lasso regression:

$$L(\mathrm{w}) = \frac{1}{2}\sum_{i=1}^{N}(f(x_i, \mathrm{w}) - y_i)^2 + \lambda \sum_{j=0}^{p} |w_j|.$$

- The main difference is that the form of the constraints is now such that it is much more probable to get strictly zero w_j.
- Btw, what do I mean by "form of the constraints"?

Lasso

- We can rewrite the regression with regularizer in a different way:

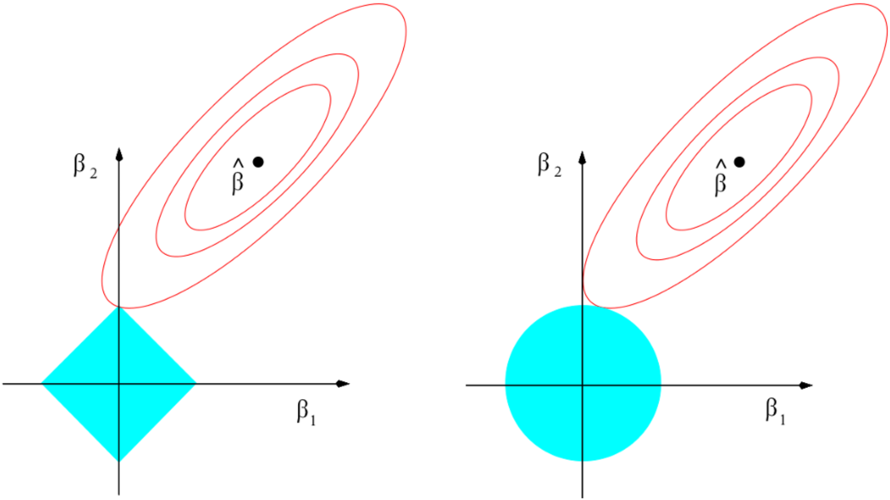
$$\mathrm{w}^* = \arg\min_{\mathrm{w}} \left\{ \frac{1}{2} \sum_{i=1}^{N} (f(x_i, \mathrm{w}) - y_i)^2 + \lambda \sum_{j=0}^{p} |w_j| \right\},$$

is equivalent to

$$\mathrm{w}^* = \arg\min_{\mathrm{w}} \left\{ \frac{1}{2} \sum_{i=1}^{N} (f(x_i, \mathrm{w}) - y_i)^2 \right\} \text{ for } \sum_{j=0}^{p} |w_j| \leq t.$$

Prove it.

Ridge and lasso



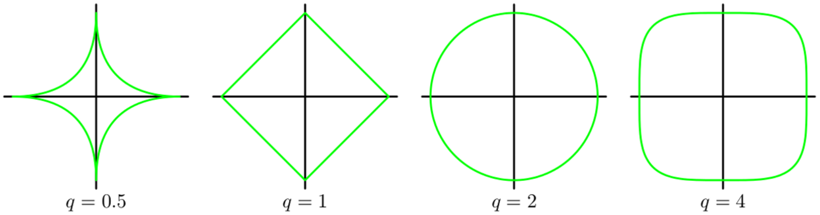
Generalization

- We can generalize ridge and lasso regression to

$$L(\mathrm{w}) = \frac{1}{2}\sum_{i=1}^{N}(f(x_i, \mathrm{w}) - y_i)^2 + \lambda \sum_{j=0}^{p}(|w_j|)^q.$$

Which prior distribution on w **does this correspond to?**

Different q



Thank you!

Thank you for your attention!