LINEAR REGRESSION

Master's Deep Learning

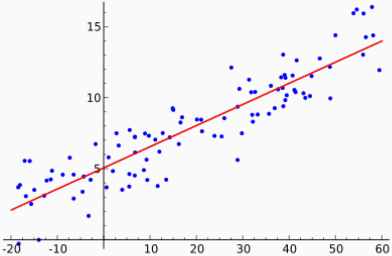
Sergey Nikolenko

Harbour Space University, Barcelona, Spain
November 8, 2017

LINEAR REGRESSION

- For example, *linear regression*.
- Linear model: consider a linear function

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



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

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

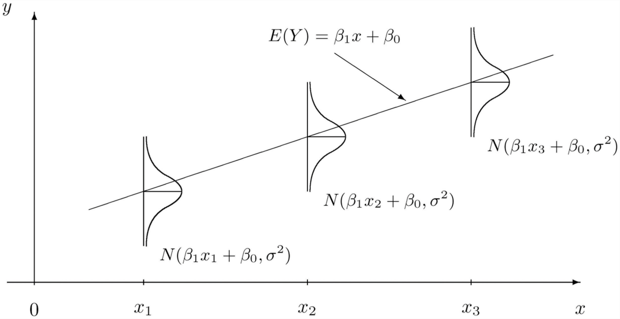
$$\mathrm{RSS}(\mathbf{w}) = \sum_{i=1}^{N} (y_i - \mathbf{x}_i^\top \mathbf{w})^2.$$

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

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

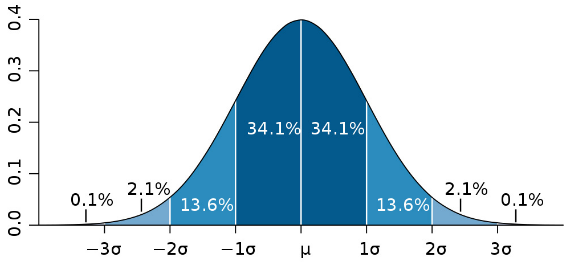
$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \text{ то есть}$$

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



- Aside – normal distribution:

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



- Why is it so important?

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

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

- We take the logarithm (we omit $\mathbf{X}$ below for brevity):

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

- And we see that to maximize the likelihood w.r.t. $\mathbf{w}$ we need to minimze mean squared error!

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

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

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

  ...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 на гипотезах/параметрах:

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

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

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

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

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mu_0, \Sigma_0).$$

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

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

- Then the problem is to compute

$$p(\mathbf{w} \mid \mathbf{t}) \propto p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \sigma^2) p(\mathbf{w})$$
$$= \mathcal{N}(\mathbf{w} \mid \mu_0, \Sigma_0) \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \mathbf{w}^\top \phi(\mathbf{x}_n), \sigma^2\right).$$

- Let us compute!

- We get

$$
\begin{aligned}
p(\mathbf{w} \mid \mathbf{t}) = \mathcal{N}(\mathbf{w} \mid \mu_N, \Sigma_N),& \\
\mu_N = \Sigma_N \left(\Sigma_0^{-1} \mu_0 + \frac{1}{\sigma^2} \Phi^\top \mathbf{t} \right),& \\
\Sigma_N = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2} \Phi^\top \Phi \right)^{-1}.&
\end{aligned}
$$

- And now the log likelihood.

- If we take the prior distribution around zero:

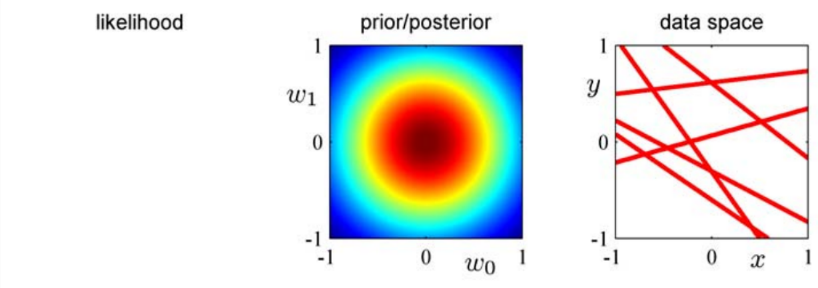
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid 0, \frac{1}{\alpha}\mathbf{I}),$$

we get the log likelihood as

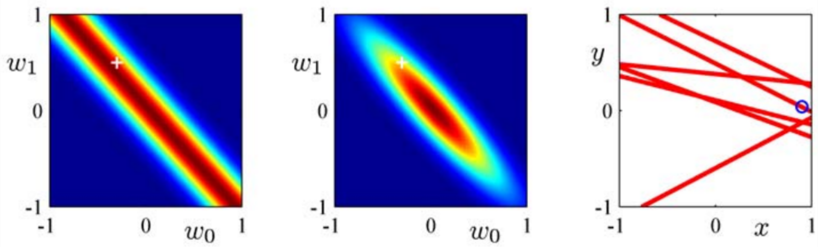
$$\ln p(\mathbf{w} \mid \mathbf{t}) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)\right)^2 - \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w} + \text{const},$$

i.e., precisely ridge regression!

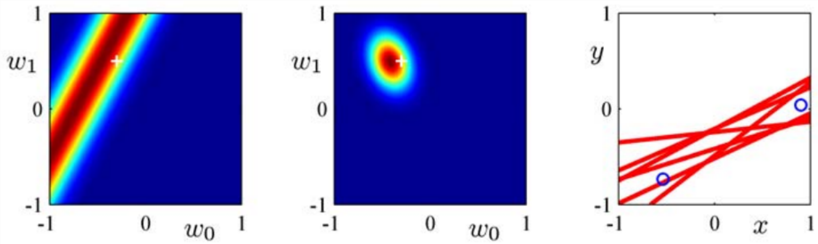
EXAMPLE



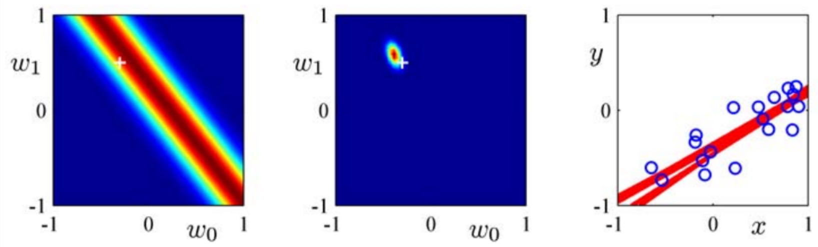
likelihood
prior/posterior
1
w_1
0
-1
-1
0
w_0
1
data space
1
y
0
-1
-1
0
x
1



1
w_1
0
-1
-1
0
w_0
1
1
w_1
0
-1
-1
0
w_0
1
1
y
0
-1
-1
0
x
1



1
w_1
0
-1
-1
0
w_0
1
1
w_1
0
-1
-1
0
w_0
1
1
y
0
-1
-1
0
x
1



1
w_1
0
-1
-1
0
w_0
1
1
w_1
0
-1
-1
0
w_0
1
1
y
0
-1
-1
0
x
1

- A slight generalization – a more general prior distribution:

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

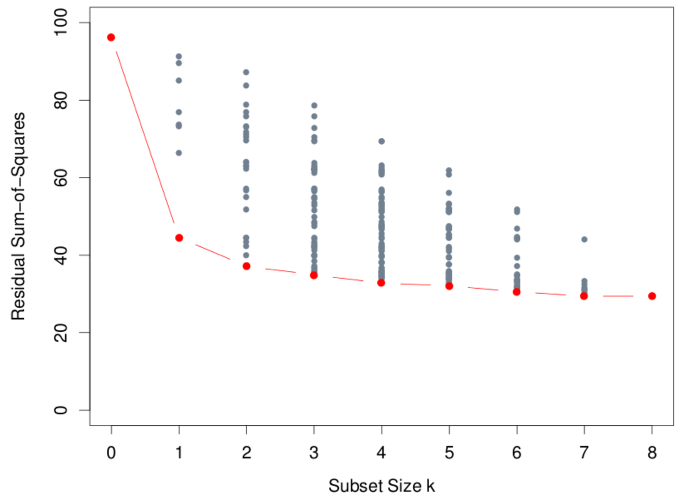
Упражнение. Compute the log likelihood.

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 $\mathbf{w}$.

- 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

- 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.


Residual Sum−of−Squares
0
20
40
60
80
100
0
1
2
3
4
5
6
7
8
Subset Size k

LASSO

- Let us now consider lasso regression:

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

- 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”?

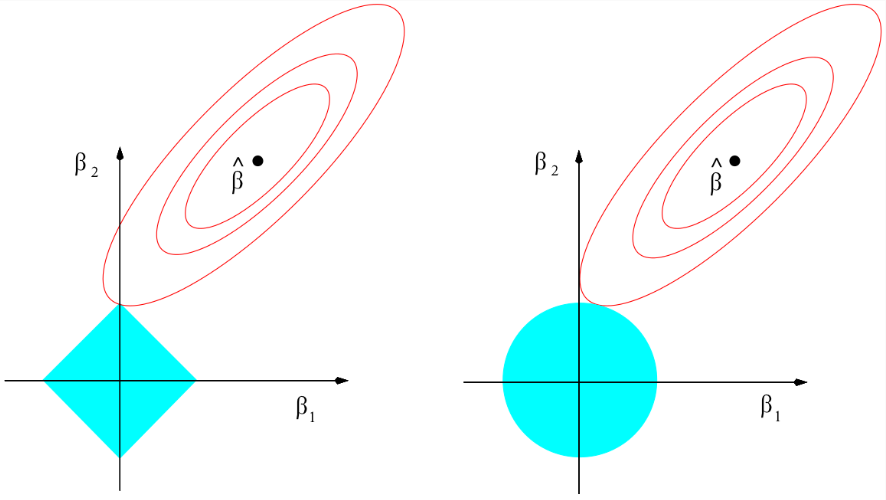
- We can rewrite the regression with regularizer in a different way:

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

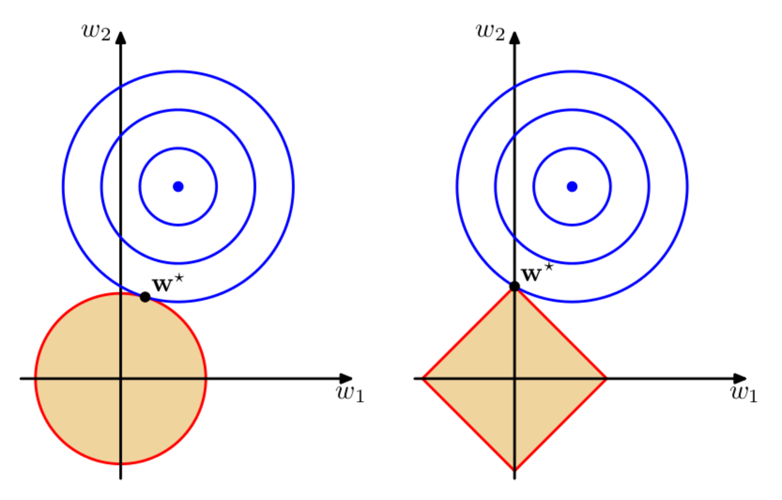
is equivalent to

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

Упражнение. Prove it.


β 2
β̂
β 1
β 2
β̂
β 1

RIDGE AND LASSO

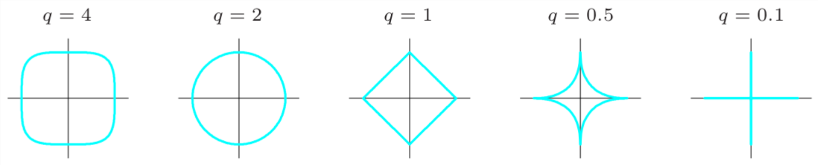


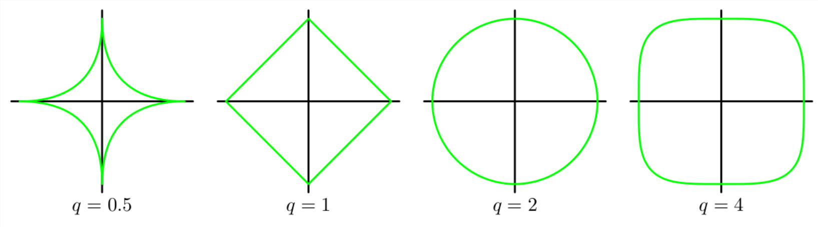
- We can generalize ridge and lasso regression to

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

Упражнение. Which prior distribution on $\mathbf{w}$ does this correspond to?

DIFFERENT q





THANK YOU!

Thank you for your attention!