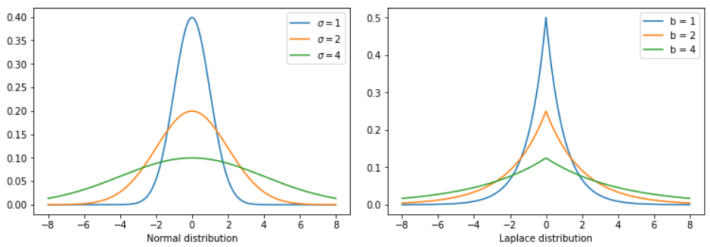
Introduction to Machine Learning

Nonlinear and Bayes



Learning goals

- Know how regularization can be motivated from a Bayesian perspective
- Understand the correspondence between log-prior and regularization term

SUMMARY: REGULARIZED RISK MINIMIZATION

In $\mathcal{R}_{\text{reg}}$ one has extreme flexibility to make appropriate choices

$$\mathcal{R}_{\text{reg}}(f) = \min_{f \in \mathcal{H}} \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right) + \lambda \cdot J(f)$$

for a given ML problem:

- the **representation** of f, which determines how features can influence the predicted y
- the **loss** function, which measures how errors should be treated
- the **regularization** $J(f)$, which encodes our inductive bias and preference for certain simpler models

By varying these choices one can construct a huge number of different ML models. Many ML models follow this construction principle or can be interpreted through the lens of regularized risk minimization.

Regularization from a Bayesian Perspective

REGULARIZED RISK MINIMIZATION VS. BAYES

We have already created a link between maximum likelihood estimation and empirical risk minimization.

Now we will generalize this for regularized risk minimization.

Assume we have a parameterized distribution $p(\mathbf{x}|\boldsymbol{\theta})$ for our data and a prior $p(\boldsymbol{\theta})$ over our parameter space, all in the Bayesian framework.

With Bayes theorem we know:

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})} \propto p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

REGULARIZED RISK MINIMIZATION VS. BAYES

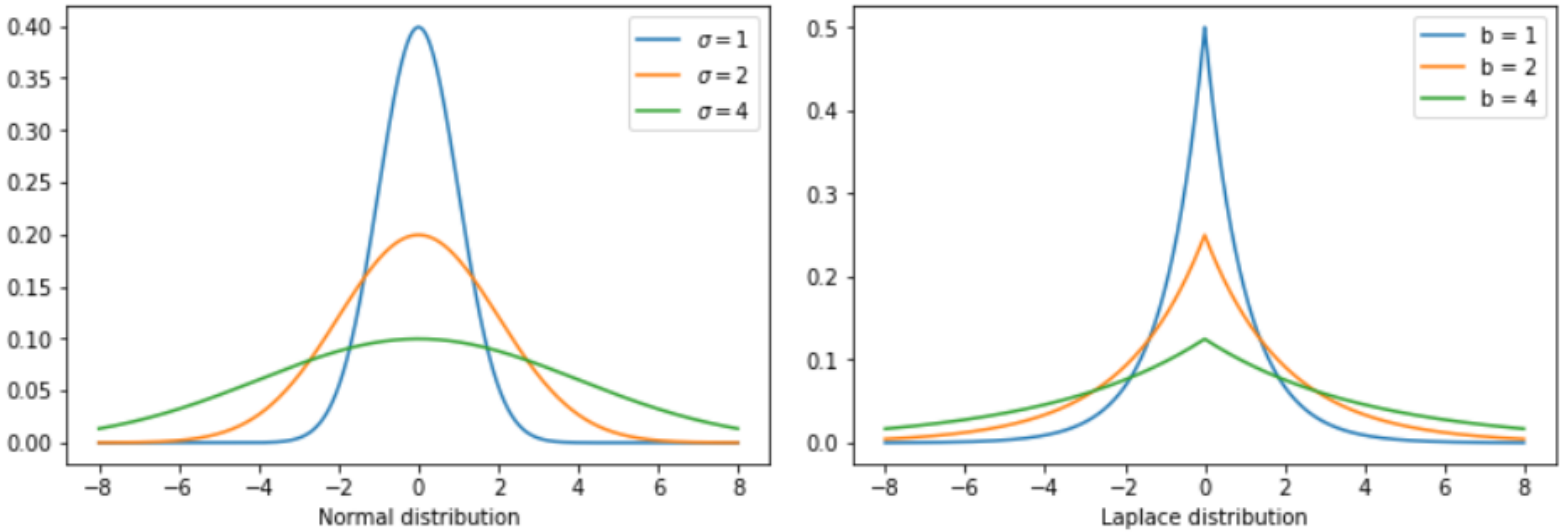
The maximum a posteriori (MAP) estimator of $\boldsymbol{\theta}$ is now the minimizer of

$$-\sum_{i=1}^{n} \log p\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right) - \log p(\boldsymbol{\theta}).$$

Again, we identify the loss $L\left(y, f(\mathbf{x} \mid \boldsymbol{\theta})\right)$ with $-\log(p(\mathbf{x}|\boldsymbol{\theta}))$. If $p(\boldsymbol{\theta})$ is constant (i.e., we used a uniform, non-informative prior), we arrive at empirical risk minimization.

If not, we can identify $J(\boldsymbol{\theta}) \propto -\log(p(\boldsymbol{\theta}))$, i.e., the log-prior corresponds to the regularizer, and the additional control parameter λ corresponds to the relative strength of the prior in regularized risk minimization.

REGULARIZED RISK MINIMIZATION VS. BAYES



- L_2 regularization corresponds to a zero-centered Gaussian prior, $\theta_i \sim \mathcal{N}(0, \sigma^2)$.
- L_1 regularization corresponds to a zero-centered Laplace prior, $\theta_i \sim \text{Laplace}(0, b) = \frac{1}{2b} \exp(-\frac{|\theta_i|}{b})$, where b is a scale parameter.
- In both cases, as regularization strength λ increases, the variance of the prior decreases, which in turn shrinks the parameters.