Reminder: the course material is at : `https://gitlab.inria.fr/flandes/fpml`

3 Regularizations

3.1 Gradients (of regularization terms)

Check that you're able to make the proof for the exact solution of linear regression, without regularization, in d dimensions (and check what it means for $d = 1$ dimensional input).

Check that you're able to make the proof for the exact solution of Ridge-regularized linear regression, i.e. add a term $+\lambda||w||_2^2$ to the loss. Remember that $||w||_2^2 = \vec{w}^T\vec{w}$.

Check the meaning of the general formula, in the special case $D = 1$, where matrix inverses are "easy" (trivial, really).

3.2 Weight shrinkage

Having computed the gradient of a Ridge-regularized linear regression (see exercise above), write down the GD update step, and how it corresponds to shrinking (geometrical decay).

3.3 Lasso regularization

Try to find the exact solution of Lasso-regularized linear regression, i.e. with a term $\lambda||w||_1 = \lambda\sum_d^D |w_d|$. Study each dimension separately, then introduce the appropriate sign() operator on vectors. What's the problem with our attempt at an exact solution?

Having computed the gradient of a Lasso-regularized linear regression, write down the GD update step, and how it corresponds to some kind of shrinking. This is a naïve gradient. The real solution is seen later (next excerise, sub-gradients).

We want to understand why these updates are making some coefficients to be exactly 0. This exercise is completed with the notebook "**toy example Lasso.ipynb**" This can be understood by plotting a toy loss, e.g. using a $D = 1$ dimensional input, and plotting the Loss, $\sum_n(wx_n - t_n)^2 + \lambda||w||_1$. Do it for a data set of a few points, such that the non-regularized optimal w^* is close to 0. Then plot the Lasso Loss for $\lambda = 1$. Then again but for a smaller λ. You should then also plot your model, $y = w.x$, and the 4 data points.

On the same plot, show the case of Ridge (L2) regression. Why doesn't it also give exactly $w = 0$ coefficients?

There is also the L_0 norm, with a generic term $||w||_0 = \{1 \text{ if } w \neq 0, \text{ else}, 0\}$, i.e. it simply counts the number of non-zero coefficients. Think about it and guess why it is aims in the same direction as Lasso, but people usually prefer to use Lasso in their algorithms.

3.4 Maximum A Posteriori (in general)

Compute the MAP estimation of a variable X that is expected to follow a Gaussian law, $\mathcal{N}(\mu, \sigma)$, where we have an exponential prior for the mean: $\mu \sim \lambda e^{-\lambda\mu}$.

Compute the MAP estimation of a variable X that is expected to follow a Gaussian law, $\mathcal{N}(\mu, \sigma)$, where we have a Laplace[1] prior for the mean: $\mu \sim Laplace(0, b) = \frac{1}{2b}e^{-\frac{|x-0|}{b}}$.

3.5 Maximum A Posteriori (for regularization)

Assume that outputs y follow a linear model perturbed by Gaussian noise: $y = \vec{w}^T\vec{x} + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Check that you know how to recover Ridge regression by assuming a Gaussian prior on the weights of the model.

Assuming a Laplace prior for each weight w_d of the model, $w_d \sim Laplace(0, b) = \frac{1}{2b}e^{-\frac{|x-0|}{b}}$, what Loss do you get?

3.6 Standardization and regularization

Is it better to standardize the input data, or not, when we do regularization? Why is that? (The Bayesian interpretation may help you). In particular, think about a dataset where each input feature has a different unit, like in the boston house market dataset (square feet, dollars, number of rooms, number of windows, etc).

Why is it probably a bad idea to use regularization on the bias (say, in regression)? If we do use such regularization, what trick can we apply to the labels to balance this issue ?

[1] Also called double exponential distribution, although this can be confused with the Gumbel distribution.