Machine Learning Homework Sheet 03

Linear Regression

We changed the deadlines for submitting homework solutions to Sundays.
Please upload the solution to Moodle by Sunday, 11.11.2018, 23:59 CET.

1 Least squares regression

Problem 1: Load the notebook `03_homework_linear_regression.ipynb` from Piazza. Fill in the missing code and run the notebook. Convert the evaluated notebook to pdf and add it to the printout of your homework.

Note: We suggest that you use Anaconda for installing Python and Jupyter, as well as for managing packages. We recommend that you use Python 3.

For more information on Jupyter notebooks and how to convert them to other formats, consult the Jupyter documentation and nbconvert documentation.

Problem 2: Let's assume we have a dataset where each datapoint, $(\boldsymbol{x}_i, y_i)$ is weighted by a scalar factor which we will call t_i. We will assume that $t_i > 0$ for all i. This makes the sum of squares error function look like the following:

$$E_{\text{weighted}}(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N} t_i \left[\boldsymbol{w}^T \boldsymbol{\Phi}(x_i) - y_i\right]^2$$

Find the equation for the value of $\boldsymbol{w}$ that minimizes this error function.

Furthermore, explain how this weighting factor, t_i, can be interpreted in terms of
1) the variance of the noise on the data and
2) data points for which there are exact copies in the dataset.

If we define $\boldsymbol{T} = \text{diag}(t_1, \ldots, t_N)$ to be a diagonal matrix containing the weighting coefficients, then we can write the weighted sum-of-squares cost function in the form

$$E_{\text{weighted}}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{\Phi w} - \boldsymbol{y})^T \boldsymbol{T}(\boldsymbol{\Phi w} - \boldsymbol{y})$$

Setting the derivative with respect to $\boldsymbol{w}$ to zero, and re-arranging, then gives

$$\boldsymbol{w}^*_{\text{weighted}} = (\boldsymbol{\Phi}^T \boldsymbol{T} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \boldsymbol{T} \boldsymbol{y}$$

which reduces to the standard solution for the case $\boldsymbol{T} = \boldsymbol{I}$. I.e.

$$\boldsymbol{w}_{\text{ML}} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}$$

If you remember back to when we modeled the likelihood using a Gaussian, our likelihood had the following form:

$$p(\boldsymbol{y} \mid \boldsymbol{\Phi}, \boldsymbol{w}, \beta) = \prod_{i=1}^{N} \mathcal{N}(y_i \mid \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_i}), \beta^{-1})$$

After applying the logarithm and using the standard form for the univariate Gaussian our equation looked like this:

$$\begin{aligned}\ln p(\boldsymbol{y} \mid \boldsymbol{\Phi}, \boldsymbol{w}, \beta) &= \sum_{i=1}^{N} \ln \mathcal{N}(y_n \mid \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i), \beta^{-1}) \\ &= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_{\text{LS}}(\boldsymbol{w})\end{aligned}$$

Where $E_{\text{LS}}(\boldsymbol{w})$ is the standard sum of squares error function (not to be confused with the $E_{\text{weighted}}(\boldsymbol{w})$ we defined earlier). Remember that $E_{\text{LS}}(\boldsymbol{w})$ is defined as follows:

$$E_{\text{LS}}(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i) - y_i)^2$$

When we compare E_{LS} with E_{weighted} and the effect of swapping the two in the previous likelihood equation we can see that T_i can be regarded as a precision (inverse variance) parameter, particular to the data point $(\boldsymbol{x}_i, y_i)$, that either replaces or scales β.

Alternatively, t_i can be regarded as an *effective* number of replicated observations of data point $(\boldsymbol{x}_i, y_i)$; this becomes particularly clear if we consider $E_{\text{weighted}}(\boldsymbol{w})$ with t_i taking positive integer values, although it is valid for any $t_i > 0$.

2 Ridge regression

Problem 3: Show that the following holds: The ridge regression estimates can be obtained by ordinary least squares regression on an augmented dataset: Augment the design matrix $\boldsymbol{\Phi} \in \mathbb{R}^{N\times M}$ with M additional rows $\sqrt{\lambda}\boldsymbol{I}_{M\times M}$ and augment $\boldsymbol{y}$ with M zeros.

Ordinary least squares minimizes $(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w})$. For ridge regression we need to minimize $(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w}) + \lambda\boldsymbol{w}^T\boldsymbol{w}$. If we define $\hat{\boldsymbol{\Phi}} = \begin{pmatrix}\boldsymbol{\Phi} \\ \sqrt{\lambda}\boldsymbol{I}\end{pmatrix}$ and $\hat{\boldsymbol{y}} = \begin{pmatrix}\boldsymbol{y} \\ \boldsymbol{0}_M\end{pmatrix}$, we can formulate the ridge regression objective as minimizing $(\hat{\boldsymbol{y}} - \hat{\boldsymbol{\Phi}}\boldsymbol{w})^T(\hat{\boldsymbol{y}} - \hat{\boldsymbol{\Phi}}\boldsymbol{w})$.

Upload a single PDF file with your solution to Moodle by 11.11.2018, 23:59 CET. We recommend to typeset your solution (using LaTeX or Word), but handwritten solutions are also accepted. If your handwritten solution is illegible, it won't be graded and you waive your right to dispute that.

3 Bayesian linear regression

In the lecture we made the assumption that we already knew the precision (inverse variance) for our Gaussian distributions. What about when we don't know the precision and we need to put a prior on that as well as our Gaussian prior that we already have on the weights of the model?

Problem 4: It turns out that the conjugate prior for the situation when we have an unknown mean and unknown precision is a normal-gamma distribution (See section 2.3.6 in Bishop). This is also true when we have a conditional Gaussian distribution of the linear regression model. This means that if our likelihood is as follows:

$$p(\boldsymbol{y} \mid \boldsymbol{\Phi}, \boldsymbol{w}, \beta) = \prod_{i=1}^{N} \mathcal{N}(y_i \mid \boldsymbol{w}^T \boldsymbol{\Phi}(x_i), \beta^{-1})$$

Then the conjugate prior for both $\boldsymbol{w}$ and β is

$$p(\boldsymbol{w}, \beta) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_0, \beta^{-1}\boldsymbol{S}_0)\text{Gamma}(\beta \mid a_0, b_0)$$

Show that the posterior distribution takes the same form as the prior, i.e.

$$p(\boldsymbol{w}, \beta \mid \mathcal{D}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_N, \beta^{-1}\boldsymbol{S}_N)\text{Gamma}(\beta \mid a_N, b_N)$$

Also be sure to give the expressions for $\boldsymbol{m}_N$, $\boldsymbol{S}_N$, a_N, and b_N.

It is easiest to work in log space. The log of the posterior distribution is given by

$$\ln p(\boldsymbol{w}, \beta \mid \mathcal{D}) = \ln p(\boldsymbol{w}, \beta) + \sum_{i=1}^{N} \ln p(y_i \mid \boldsymbol{w}^T \phi(\boldsymbol{x}_i), \beta^{-1})$$

$$= \frac{M}{2} \ln \beta - \frac{1}{2} \ln |\boldsymbol{S}_0| - \frac{\beta}{2} (\boldsymbol{w} - \boldsymbol{m}_0)^T \boldsymbol{S}_0^{-1} (\boldsymbol{w} - \boldsymbol{m}_0) - b_0 \beta + (a_0 - 1) \ln \beta$$

$$+ \frac{N}{2} \ln \beta - \frac{\beta}{2} \sum_{i=1}^{N} \{\boldsymbol{w}^T \phi(\boldsymbol{x}_i) - y_i\}^2 + const$$

Using the product rule, the posterior distribution can be written as $p(\boldsymbol{w}, \beta \mid \mathcal{D}) = p(\boldsymbol{w} \mid \beta, \mathcal{D}) p(\beta \mid \mathcal{D})$. Consider first the dependence on $\boldsymbol{w}$. We have

$$\ln p(\boldsymbol{w} \mid \beta, \mathcal{D}) = \frac{-\beta}{2} \boldsymbol{w}^T [\boldsymbol{\Phi}^T \boldsymbol{\Phi} + \boldsymbol{S}_0^{-1}] \boldsymbol{w} + \boldsymbol{w}^T [\beta \boldsymbol{S}_0^{-1} \boldsymbol{m}_0 + \beta \boldsymbol{\Phi}^T \boldsymbol{y}] + const$$

Thus we see that $p(\boldsymbol{w} \mid \beta, \mathcal{D})$ is a Gaussian distribution with mean and covariance given by

$$\boldsymbol{m}_N = \boldsymbol{S}_N[\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \boldsymbol{\Phi}^T\boldsymbol{y}]$$
$$\boldsymbol{S}_N^{-1} = (\boldsymbol{S}_0^{-1} + \boldsymbol{\Phi}^T\boldsymbol{\Phi})$$

To find $p(\beta \mid \mathcal{D})$ we first need to complete the square over $\boldsymbol{w}$ to ensure that we pick up all terms involving β (any terms independent of β may be discarded since these will be absorbed into the normalization coefficient which itself will be found by inspection at the end). We also need to remember that a factor of $(M/2)\ln\beta$ will be absorbed by the normalization factor of $p(\boldsymbol{w} \mid \beta, \mathcal{D})$. Thus

$$\ln p(\beta|\mathcal{D}) = \frac{-\beta}{2}\boldsymbol{m}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \frac{\beta}{2}\boldsymbol{m}_N^T\boldsymbol{S}_N^{-1}\boldsymbol{m}_N + \frac{N}{2}\ln\beta - b_0\beta + (a_0 - 1)\ln\beta - \frac{\beta}{2}\sum_{i=1}^{N} y_i^2 + const.$$

We recognize this as the log of a Gamma distribution. Reading off the coefficients of β and $ln\beta$ we then have

$$a_N = a_0 + \frac{N}{2}$$
$$b_N = b_0 + \frac{1}{2}(\boldsymbol{m}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 - \boldsymbol{m}_N^T\boldsymbol{S}_N^{-1}\boldsymbol{m}_N + \sum_{i=1}^{N} y_i^2)$$