Machine Learning Worksheet 05

Linear Regression and Kernels

1 Sum of Squared Errors Regression

Problem 1: Let's assume we have a dataset where each datapoint, Z_n is weighted by a scalar factor which we will call T_n. We will assume that $T_n > 0$. This makes the sum of squared error function look like the following:

$$E_{\mathcal{D}}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N} T_n \left[W^T\phi(x_n) - Z_n\right]^2$$

Find the equation for the value of W that minimizes this error function.

Furthermore, explain what this weighting factor, T_n, does to the error function 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 $T = \text{diag}(t_1, \ldots, t_N)$ to be a diagonal matrix containing the weighting coefficients, then we can write the weighted sum-of-squares cost function in the form

$$E_{\mathcal{D}}(W) = \frac{1}{2}(Z - \mathbf{\Phi}W)^T T (Z - \mathbf{\Phi}W)$$

.

Setting the derivative with respect to W to zero, and re-arranging, then gives

$$W^* = (\mathbf{\Phi}^T T \mathbf{\Phi})^{-1}\mathbf{\Phi}^T T Z$$

which reduces to the standard solution for the case $T = I$. I.e.

$$W_{\text{ML}} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T Z$$

.

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

$$p(Z \mid \mathbf{X}, W, \beta) = \prod_{n=1}^{N} \mathcal{N}(Z_n \mid W^T\phi(X_n), \beta^{-1})$$

.

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

$$\ln p(Z \mid \mathbf{X}, W, \beta) = \sum_{n=1}^{N} \ln \mathcal{N}(Z_n \mid W^T \phi(X_n), \beta^{-1})$$
$$= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta \hat{E_{\mathcal{D}}}(W)$$

.

Where $\hat{E_{\mathcal{D}}}(W)$ is the sum of squares error function denoted with a hat so that it isn't confused with the $E_{\mathcal{D}}(W)$ we defined earlier. Remember that $\hat{E_{\mathcal{D}}}(W)$ is defined as follows:

$$\hat{E_{\mathcal{D}}}(W) = \frac{1}{2} \sum_{n=1}^{N} \{Z_n - W^T \phi(X_n)\}^2$$

.

When we compare this version of $\hat{E_{\mathcal{D}}}$ with $E_{\mathcal{D}}$ and the effect of swapping the two in the previous likelihood equation we can see that T_n can be regarded as a precision (inverse variance) parameter, particular to the data point $(\mathbf{X}_N, Z_n)$, that either replaces or scales β.

Alternatively, T_n can be regarded as an *effective* number of replicated observations of data point $(\mathbf{X}_n, Z_n)$; this becomes particularly clear if we consider $E_{\mathcal{D}}(W)$ with T_n taking positive integer values, although it is valid for any $T_n > 0$.

2 Ridge regression

Problem 2: 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 $\boldsymbol{\Phi}$ with p additional rows $\sqrt{\lambda}\boldsymbol{I}$ and augment $\boldsymbol{z}$ with p zeros.

Ordinary least squares minimizes $(\boldsymbol{z} - \boldsymbol{\Phi w})^T(\boldsymbol{z} - \boldsymbol{\Phi w})$, i.e. solves $\boldsymbol{\Phi w} = \boldsymbol{z}$. For ridge regression we need to minimize $(\boldsymbol{z} - \boldsymbol{\Phi w})^T(\boldsymbol{z} - \boldsymbol{\Phi w}) + \lambda \boldsymbol{w}^T \boldsymbol{w}$. If we define $\hat{\boldsymbol{\Phi}} = \begin{pmatrix} \boldsymbol{\Phi} \\ \sqrt{\lambda}\boldsymbol{I} \end{pmatrix}$ and $\hat{\boldsymbol{z}} = \begin{pmatrix} \boldsymbol{z} \\ \mathbf{0}_p \end{pmatrix}$, we can formulate the ridge regression objective as minimizing $(\hat{\boldsymbol{z}} - \hat{\boldsymbol{\Phi}}\boldsymbol{w})^T(\hat{\boldsymbol{z}} - \hat{\boldsymbol{\Phi}}\boldsymbol{w})$, i.e. solving $\hat{\boldsymbol{\Phi}}\boldsymbol{w} = \hat{\boldsymbol{z}}$. p is hereby the dimension of $\boldsymbol{w}$.

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 3: 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:

$$p(Z \mid \mathbf{X}, W, \beta) = \prod_{n=1}^{N} \mathcal{N}(Z_n \mid W^T \phi(X_n), \beta^{-1})$$

Then the conjugate prior for both W and β is

$$p(W, \beta) = \mathcal{N}(W \mid M_0, \beta^{-1}\mathbf{S}_0) Gam(\beta \mid a_0, b_0)$$

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

$$p(W, \beta \mid Z) = \mathcal{N}(W \mid M_N, \beta^{-1}\mathbf{S}_N) Gam(\beta \mid a_N, b_N)$$

.

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

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

$$\ln p(W, \beta \mid Z) = \ln p(W, \beta) + \sum_{n=1}^{N} \ln p(Z_n \mid W^T \phi(X_n), \beta^{-1})$$

$$= \frac{M}{2} \ln \beta - \frac{1}{2} \ln |\mathbf{S}_0| - \frac{\beta}{2}(W - M_0)^T \mathbf{S}_0^{-1}(W - M_0) - b_0\beta + (a_0 - 1) \ln \beta$$

$$+ \frac{N}{2} \ln \beta - \frac{\beta}{2} \sum_{n=1}^{N} \{W^T \phi(X_n) - Z_n\}^2 + const$$

Using the product rule, the posterior distribution can be written as $p(W, \beta \mid Z) = p(W \mid \beta, Z) p(\beta \mid Z)$. Consider first the dependence on W. We have

$$\ln p(W \mid \beta, Z) = \frac{-\beta}{2} W^T [\mathbf{\Phi}^T \mathbf{\Phi} + \mathbf{S}_0^{-1}] W + W^T [\beta \mathbf{S}_0^{-1} M_0 + \beta \mathbf{\Phi}^T Z] + const$$

Thus we see that $p(W \mid \beta, Z)$ is a gaussian distribution with mean and covariance given by

$$M_N = \mathbf{S}_N [\mathbf{S}_0^{-1} M_0 + \mathbf{\Phi}^T Z]$$
$$\mathbf{S}_N^{-1} = \beta(\mathbf{S}_0^{-1} + \mathbf{\Phi}^T \mathbf{\Phi})$$

.

To find $p(\beta \mid Z)$ we first need to complete the square over 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 normalisation factor of $p(W \mid \beta, Z)$. Thus

$$\ln p(\beta|Z) = \frac{-\beta}{2} M_0^T \mathbf{S}_0^{-1} M_0 + \frac{\beta}{2} M_N^T \mathbf{S}_N^{-1} M_N + \frac{N}{2}\ln\beta - b_0\beta + (a_0 - 1)\ln\beta - \frac{\beta}{2}\sum_{n=1}^{N} Z_n^2 + const.$$

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

$$a_N = a_0 + \frac{N}{2}$$

$$b_N = b_0 + \frac{1}{2}(M_0^T \mathbf{S}_0^{-1} M_0 - M_N^T \mathbf{S}_N^{-1} M_N + \sum_{n=1}^{N} Z_n^2)$$

4 Facebook advertisements

You want to boost your Facebook page and therefore you book Facebook advertisements. A simple linear model for the number of new likes per week (y), depending on the money spent (x) could be:

$$\begin{aligned} y &= a_0 + a_1 x + \epsilon \\ \text{where } y &= \text{number of new likes per week} \\ x &= \text{money spent in that week, in units of 1 EUR} \\ \epsilon &= \text{normal (Gaussian) distributed fluctuations} \end{aligned}$$

After taking a lot of measurement data you fit the parameters. You find:

$$\begin{aligned} a_0 &= 10 \\ a_1 &= 5 \\ \mathbb{E}[\epsilon] &= 0 \\ \mathrm{Var}[\epsilon] &= 4 \end{aligned}$$

The full model is therefore given by

$$\begin{aligned} y &= 10 + 5x + \mathcal{N}(0, 4) \\ &\Rightarrow y \sim \mathcal{N}(5x + 10, 4) \end{aligned}$$

Problem 4: Assume you spend no money, what is the probability that you get more than 10 likes per week?

$$\begin{aligned} x &= 0 \\ \Rightarrow y &= 10 + \epsilon \longrightarrow \mathcal{N}(10, 4) \\ \Rightarrow p(y > 10) &= \int_{10}^{\infty} \mathcal{N}(10, 4) dy \\ &= 0.5 \end{aligned}$$

Problem 5: Now you spend 1 EUR on advertisements. What is the expected value of likes?

$$\begin{aligned} x &= 1 \\ \Rightarrow y &= 10 + 5 + \epsilon \longrightarrow \mathcal{N}(15, 4) \\ \Rightarrow \mathrm{E}[y] &= 15 \end{aligned}$$

5 Kernelised k-nearest neighbours

To classify the point $\vec{x}$ the k-nearest neighbours finds the k training samples $\mathcal{N} = \{\vec{x}^{(s_1)}, \vec{x}^{(s_2)}, \ldots, \vec{x}^{(s_k)}\}$ that have the shortest distance $||\vec{x} - \vec{x}^{(s_i)}||_2$ to $\vec{x}$. Then the label that is mostly represented in the neighbour set $\mathcal{N}$ is assigned to $\vec{x}$.

Problem 6: Formulate the k-nearest neighbours algorithm in feature space by introducing the feature map $\vec{\phi}(\vec{x})$. Then rewrite the k-nearest neighbours algorithm so that it only depends on the scalar product in feature space $K(\vec{x}, \vec{y}) = \vec{\phi}(\vec{x})^T \vec{\phi}(\vec{y})$.

The distance to a training sample in feature space is given by

$$|\vec{\phi}(\vec{x}) - \vec{\phi}(\vec{x}^{(s_i)})|_2 \,.$$

We can replace this by the squared distance because this will not change which points are nearest to $\vec{x}$. Thus we have

$$\begin{aligned} |\vec{\phi}(\vec{x}) - \vec{\phi}(\vec{x}^{(s_i)})|_2^2 &= (\vec{\phi}(\vec{x}) - \vec{\phi}(\vec{x}^{(s_i)}))^T (\vec{\phi}(\vec{x}) - \vec{\phi}(\vec{x}^{(s_i)})) \\ &= \vec{\phi}(\vec{x})^T \vec{\phi}(\vec{x}) - 2\vec{\phi}(\vec{x})^T \vec{\phi}(\vec{x}^{(s_i)}) + \vec{\phi}(\vec{x}^{(s_i)})^T \vec{\phi}(\vec{x}^{(s_i)}) \,. \end{aligned}$$

The first term is a constant when searching for the k training samples that minimise this function. Hence we can drop the first term and must find the k training samples $\vec{x}^{(s_i)}$ that minimise

$$\vec{\phi}(\vec{x}^{(s_i)})^T \vec{\phi}(\vec{x}^{(s_i)}) - 2\vec{\phi}(\vec{x})^T \vec{\phi}(\vec{x}^{(s_i)}) = K(\vec{x}^{(s_i)}, \vec{x}^{(s_i)}) - 2K(\vec{x}, \vec{x}^{(s_i)}) \,.$$

6 Radial Basis Kernels

We have a set of rules that we discussed in the lecture for proving that a kernel is a valid kernel.

Problem 7: Use these rules to show why a radial basis kernel

$$K(x_i, x_j) = exp\{-\frac{1}{2}\|x_i - x_j\|^2\}$$

is a valid kernel.

Hint: You might want to start by proving $k(x_i, x_j) = f(x_i)k_1(x_i, x_j)f(x_j)$ where k_1 is some other valid kernel.

$$exp\{-\frac{1}{2}\|x_i - x_j\|^2\} = exp\{-\frac{1}{2}x_i^T x_i + x_i^T x_j - \frac{1}{2}x_j^T x_j\}$$
$$= \underbrace{exp\{-\frac{1}{2}x_i^T x_i\}}_{f(x_i)} \cdot exp\{x_i^T x_j\} \cdot \underbrace{exp\{-\frac{1}{2}x_j^T x_j\}}_{f(x_j)}$$

Here $exp\{x_i^T x_j\}$ is a kernel by rule 4. Rule 3 allows us to incorporate $f(x_i)$ and $f(x_j)$.