Supervised Learning Coursework 2
Due: Noon, December 31

John Shawe-Taylor Andrew McDonald Dimitris Stamos

November, 2016

Abstract

Using a series of synthetic examples, in this exercise session you will acquaint yourselves with linear regression, Gaussian process regression and methods derived from the GP perspective.
Keywords: least squares regression (LSR), Gaussian process regression (GPR) — regularization, — training set, test set.

Report

Please prepare a report summarising your results (with at most 1 A4 page per Exercise) to be submitted electronically by 12.00 on December 31, 2016. Include plots and tables of results as necessary but you *do not* need to include code in your response, just a summary of the results together with answers to any of the questions. You may use Matlab or whichever language you prefer to prepare the plots. Best marks will be awarded for clarity of explanation. You may work in groups of at most two, but each participant should submit the same coursework with a clear reference to the name of your partner.

When multiple trials are requested, generate a new dataset for each trial. Unless otherwise specified, we measure the prediction error on a set of n points as

$$\frac{1}{n}\sum_{i=1}^{n}(y_{true,i} - y_{pred,i})^2 = \frac{1}{n}\|y_{true} - y_{pred}\|_2^2.$$

Exercise 1 (Linear Example) *Generate dataset X of 100 points x_i of 10 dimensions, and whose components are unit variance Gaussian centered at the origin, and a 10-dimensional weight vector w from the same distribution. Generate the observations y_i as $\langle x_i, w\rangle + \epsilon_i$, where the noise ϵ_i is Gaussian with standard deviation* 0.1*.*

a) Split the data into 80 training and 20 testing points. Learn the weight vector using Ridge Regression in the primal form on the training data with the correct regularisation parameter $\sigma^2 = 0.01$. Report the test error, and the difference between the learned and the true weight vectors.

b) *Apply Kernel Ridge Regression using the linear kernel ($K(x,z) = \langle x, z\rangle$) and compare the results with a). Report the distance (norm) between the observations, the predictions in a) (primal) and the predictions in b) (dual).*

Exercise 2 (Non-linear kernel) *We generate a non-linear dataset of 100 points with the quadratic kernel by exploiting the explicit representation from the notes (cf. slide 12). Given n points $x_i \in \mathbb{R}^d$ with standard Gaussian entries, generate the intermediate points $z_i = vec(x_i x_i') \in \mathbb{R}^{d^2}$ and a weight vector $w \in \mathbb{R}^{d^2}$ with standard Gaussian entries. The observations y_i are generated as $\langle z_i, w\rangle + \varepsilon_i$, where the noise ε_i is Gaussian with standard deviation σ. Due to the feature map representation the y_i are distributed quadratically with respect to the inputs x_i.*

a) *Provide a (3d) plot of the data for $d = 2$ with zero noise and verify visually that a quadratic relationship holds. You will need to appropriately choose a number of points and a random seed to make a visually convincing plot.*

b) *For $d = 20$, $n = 500$, $\sigma = 5$, and a train/test split of 25/75, apply non-linear kernel Ridge Regression with a quadratic kernel, and regularization parameter λ determined according to the correct noise level (cf. slide 20) and report the test error.*

c) *For the dataset from b), perform validation with a train/validation/test split of 25/15/60 to choose λ in the range $a\sigma^2$, for $a \in \{10^{-4}, 10^{-3.75}, \ldots, 10^{3.75}, 10^4\}$. Report on the mean test error and mean λ chosen by selecting based on the validation over 100 trials.*

Exercise 3 (Error bars and fitting) *We generate a dataset for the Gaussian kernel using a specified covariance (see Gaussian Process supplement on Moodle). Generate inputs $X \in \mathbb{R}^{n\times d}$ with standard Gaussian entries. Compute the kernel matrix K_{true} over all of the inputs using a radial basis function (Gaussian) kernel with width ν, that is $K_{true}(x,z) = \exp\left(-\frac{\|x-z\|^2}{2\nu^2}\right)$. Compute the Cholesky decomposition $K_{true} = LL'$. The observation vector is then generated as $y = Lu + \varepsilon$, where $u \in \mathbb{R}^n$ has standard Gaussian entries and the components of ε are Gaussian with standard deviation σ. Finally, split X and y into test sets.*

a) *Provide a (3d) plot of the data for $d = 2$ with zero noise and verify visually the type of function that is generated. You will need to appropriately choose a number of points, kernel parameter and random seed to make a visually convincing plot.*

b) *Set $d = 10$, $n = 500$, $\nu^0 = 5$, $\sigma = 1.3$ and use a train/validation/test split of 12.5/12.5/75. Set the regularization parameter λ to the true value (σ^2). Using non-linear kernel Ridge Regression with a radial basis function (Gaussian) kernel width parameter ν, use the fit of the predicted error*

bar vs. a standard Gaussian to select the value of the width by validating (minimizing) the distance from the empirical cumulative distribution function (CDF) of the normalized errors to a Gaussian CDF over a range $a\nu^0$, for $a \in \{10^{-2}, 10^{-1.5}, \ldots, 10^{1.5}, 10^2\}$. Note that during the validation phase the validation set plays the role of the test set. Report on the mean test error and mean ν chosen by validation over 100 trials.

Hint: *Validating using the predicted error bars.*

Compute the cumulative distribution of the normalised predicted errors, compare to that of a Gaussian distribution, validate over the minimum distance between the two.

Explicitly, for test point x_i with given prediction $y_{pred,i}$ obtained by kernel Ridge Regression using the kernel K over the training data:

- *compute the predicted error variance (cf. slide 30)*

$$\sigma_i^2 = \kappa(x_i, x_i) - k_i'(K + \lambda I)^{-1} k_i$$

  where κ is kernel matrix, k_i is the (column) vector formed by the values $\kappa(x_{train,j}, x_i)$ for $j \in \{1, \ldots, n_{train}\}$, and K is the kernel matrix over the training data

- *compute the normalised errors $err_i = (y_{true,i} - y_{pred,i})/\sigma_i$*

- *compute the histogram of the errors (e.g. with 50 buckets), normalized by the number of test points (to ensure the sum is 1)*

- *compute the cumulative sum of this distribution to produce the empirical CDF and compute its distance to that of a standard Gaussian*

Exercise 4 (Evidence based model selection) *Generate a dataset using the same method as in Exercise 3, with $d = 10$, $n = 100$, $\nu = 5$, $\sigma = 0.5$ and use a train/validation/test split of* 12.5/12.5/75.

a) *Using non-linear kernel Ridge Regression with a radial basis function (Gaussian) kernel with width equal to the true width ν, and use the evidence formula (cf. slide 31) to select the value of the regularization parameter λ by validating (maximizing) over a range $a\sigma^2$, for $a \in \{10^{-2}, 10^{-1.9}, \ldots, 10^{1.9}, 10^2\}$. Note that the evidence is computed on the training data and the validation set is not used. Report on the mean test error and mean λ chosen over 100 trials.*

b) *Perform the model selection using the error on the validation set and compare performance (test error) with that of a), again over 100 trials.*