Exercises
Machine Learning
AS 2012

Machine Learning Laboratory
Dept. of Computer Science, ETH Zürich
Prof. Dr. Joachim M. Buhmann
Web http://ml2.inf.ethz.ch/courses/ml/

Series 6, Dec 4th, 2012 (Regression & Parzen Windows)

Email questions to: Dwarikanath Mahapatra
dwarikanath.mahapatra@inf.ethz.ch

Problem 1 (Non-Linear Regression using Basis Functions):

In this problem, we consider a way of extending the linear regression technique to non-linear problems. A simple (but often effective) approach is linear regression with basis expansion. Instead of actually fitting a non-linear function to the data, the data is preprocessed by a non-linear mapping, and linear regression is then applied to the transformed problem. This method is equivalent to fitting a regression function of the form

$$\hat{f}(x) = \sum_{j=0}^{d} \beta_j h_j(x) \ , \tag{1}$$

where the *basis functions* h_j are fixed. The linear coefficients β_j are the target parameters of the learning problem. β_0 accounts for the bias, and the corresponding basis function is the constant function $h_0(x) = 1$.

To apply linear regression in the one-dimensional case, we transform each observation $x^{(i)}$ into a vector $\mathbf{h}(x^{(i)}) := \big(h_0(x^{(i)}), \ldots, h_d(x^{(i)})\big)$. The input data vector $\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})^\top$ is then substituted by a matrix $\mathbf{H}$, containing the vectors $\mathbf{h}(x^{(i)})$ as its rows:

$$\begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(n)} \end{bmatrix} \text{becomes} \begin{bmatrix} h_0(x^{(1)}) & h_1(x^{(1)}) & \ldots & h_d(x^{(1)}) \\ h_0(x^{(2)}) & h_1(x^{(2)}) & \ldots & h_d(x^{(2)}) \\ \vdots & \vdots & \ddots & \vdots \\ h_0(x^{(n)}) & h_1(x^{(n)}) & \ldots & h_d(x^{(n)}) \end{bmatrix} .$$

Both standard linear regression and ridge regression are now applicable.

Here is what we would ask you to do:

1. Generate values from a sinc function $f(x) = \text{sinc}(3x)$, by uniformly sampling the input domain and adding zero-mean Gaussian noise to the function values, to obtain:

$$y = f(x) + \epsilon \ , \tag{2}$$

   where $x \sim \text{Uniform}(0, 1)$ and $\epsilon \sim N(0, 0.01)$. The input vector $\mathbf{x}$ consists of $n = 25$ samples and the corresponding output vector is $\mathbf{y}$.

2. Program a regression estimator. The basis functions we consider are Gaussians of the form.

$$h_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right) \tag{3}$$

   for $j = 1, \ldots, d$. Program an estimation routine

```
beta = regress_gauss(x,y,means,var,lambda)
```

   The arguments are:

   - `x` - a single observation vector
   - `y` - the corresponding outputs
   - `means` - vector of basis function means μ_j
   - `var` - the (scalar) variance parameter of the basis functions
   - `lambda` - the ridge regression regularization parameter

- `beta` - the estimate $\hat{\beta}$ or $\hat{\beta}^{\text{ridge}}$, respectively.

3. Apply your estimator to the data you generated. Use the following values for the parameters:
   - $d = 21$ (a total of 22 functions including the bias)
   - μ_j equally spaced over the interval $[0, 1]$.
   - $\sigma_j = 0.04$
   - $\lambda \in \{0.001, 0.05, 5\}$.

   For each λ, plot a graph that shows:

   - the function estimated from training vector $\mathbf{x}$
   - the function estimated by averaging over the $\boldsymbol{\beta}^{(i)}$ obtained from repeating the above experiment 100 times, i.e. generating pairs $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$, $i = 1, \ldots, 100$, estimating the corresponding $\boldsymbol{\beta}^{(i)}$, and averaging over all runs to obtain $\boldsymbol{\beta}^{avg}$.

   The estimated function is given by:

   $$\hat{f}(x) = \beta_0^{\text{ridge}} + \sum_{j=1}^{d} h_j(x)\beta_j^{\text{ridge}} . \tag{4}$$

   To plot the function, compute the $\hat{f}(x)$ values for $x \in [0, 1]$ equally spaced (e.g. `linspace` function). Comment about the plots and the quality of solutions.

4. The estimation error can be computed by looking at the squared error of the estimator to the true function. Please plot the average error as a function of λ. To this end, repeat the estimation process for the 100 random vectors $\mathbf{x}^{(i)}$, $i = 1, \ldots, 100$, for the regularizations `lambda = exp([-5:0.3:2])`.

   Instead of computing the L_2 distance between the functions by integration, we approximate it by

   $$err = \frac{1}{m}\sum_{i=1}^{m}(f(z_i) - \hat{f}(z_i))^2 , \tag{5}$$

   where $z_1, \ldots, z_m$ are equally spaced points ($m = 100$). Plot the average errors to show how it varies over λ. Comment on what happens on the two ends of the plot.

Problem 2 (Shift of the Eigenvalue Spectrum):

In this problem, analyzing the regularization mechanism of ridge regression will serve as an excuse to do some linear algebra.

1. Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric matrices. Assume that $\xi \in \mathbb{R}^n$ is an eigenvector for *both* matrices, with eigenvalues λ_A, λ_B respectively. Please show that ξ is an eigenvector of $A + B$. What is the corresponding eigenvalue?

2. Now consider the ridge regression solution

   $$\widehat{\beta}^{ridge} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top y.$$

   Please use the result in 1.) to explain why computing $\hat{\beta}^{\text{ridge}}$ is more stable than $\hat{\beta}$, i.e. why the solution can be reliably computed (for a suitable λ) even if $\mathbf{X}^\top\mathbf{X}$ is numerically singular.

3. Compare linear regression and ridge regression in terms of the bias-variance trade-off.

Problem 3 (Parzen windows):

Let $p(x) \sim U(0, a)$ be uniform from 0 to a, and let a Parzen window be defined as $\phi(x) = e^{-x}$ for $x > 0$ and 0 for $x \leq 0$.

1. Show that the mean of such a Parzen-window estimate is given by:

$$\bar{p}_n(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{a}(1 - e^{-x/h_n}) & 0 \leq x \leq a \\ \frac{1}{a}(e^{a/h_n} - 1)e^{-x/h_n} & a \leq x \end{cases} \tag{6}$$

2. Plot $\bar{p}_n(x)$ versus x for $a = 1$ and $h_n = 1, 0.25$ and 0.0625.
3. How small does h_n have to be to have less than 1% bias over 99% of the range $0 < x < a$?
4. Find h_n for this condition if $a = 1$, and plot $\bar{p}_n(x)$ in the range $0 \leq x \leq 0.05$.