Exercises for
Pattern Recognition
Peter Fischer, Shiyang Hu
Assignment 7, 2./5.12.2014

SOLUTION

General Information:

Exercises (1 SWS):	Tue 12:15 – 13:45 (0.154-115) and Fri 08:15 – 09:45 (0.151-115)
Certificate:	Oral exam at the end of the semester
Contact:	peter.fischer@fau.de, shiyang.hu@fau.de

Regression

Exercise 1 The goal of this exercise is robust regression line fitting for N measurements (x_i, y_i). Thus, you should estimate parameters a, b for a line $ax_i + b$ that best explains your observations y_i. Here we employ the Huber norm to make the estimate more robust to outliers compared to simple least-square regression:

$$(a, b) = \arg\min_{a,b} D(a, b) = \arg\min_{a,b} \sum_{i=1}^{N} \phi_{\text{Huber}}\left(y_i - ax_i - b\right) \tag{1}$$

The parameters (a, b) are determined using iterative numerical optimization. The Huber norm is defined as

$$\phi_{\text{Huber}}(z) = \begin{cases} z^2 & \text{if } |z| \leq M \\ M(2|z| - M) & \text{if } |z| > M \end{cases} \tag{2}$$

(a) Calculate the gradient of the cost function w.r.t. a and b. The gradient is necessary for many iterative numerical optimization techniques.
Hint: You need to calculate the derivative of the Huber norm.

Derivative of the Huber norm:

$$\frac{\partial \phi_{\text{Huber}}(z)}{\partial z} = \begin{cases} 2z & \text{if } |z| \leq M \\ 2M & \text{if } z > M \\ -2M & \text{if } z < -M \end{cases} \tag{3}$$

Derivative of the cost function w.r.t. a:

$$\frac{\partial D(a, b)}{\partial a} = \sum_{i=1}^{N} \phi'_{\text{Huber}}\left(y_i - ax_i - b\right)(-x_i)$$

$$\frac{\partial D(a, b)}{\partial b} = \sum_{i=1}^{N} \phi'_{\text{Huber}}\left(y_i - ax_i - b\right)(-1)$$

(b) Show that the Huber norm is convex. Use the first-order convexity condition for differentiable functions $f(x)$

$$f(z) \geq f(x) + f'(x)(z - x)$$

Start by proving convexity for $g(x) = x^2$ and $h(x) = M(2|x| - M)$. Then, treat the special cases that occur due to the piece-wise definition of the Huber norm. For this exercise, focus only on positive values x, z, M.

The domain of the function is $\mathbb{R}$, which is a convex set.
Convexity of $g(x) = x^2$:

$$\begin{aligned} z^2 &\overset{!}{\geq} x^2 + 2x(z-x) = x^2 + 2xz - 2x^2 = 2xz - x^2 \\ &= z^2 - z^2 + 2xz - x^2 = z^2 - \underbrace{(z-x)^2}_{\geq 0} \end{aligned}$$

Convexity of $M(2|x| - M)$, shown only for $x, z > 0$:

$$\begin{aligned} M(2z - M) &\overset{!}{\geq} M(2x - M) + 2M(z - x) \\ 2z - M &\overset{!}{\geq} 2x - M + 2z - 2x \\ 2z - M &\geq 2z - M \end{aligned}$$

The Huber function involves two special cases due to the piece-wise definition.
Case 1: $|x| > M$, but $|z| \leq M$

$$\begin{aligned} z^2 &\overset{!}{\geq} M(2x - M) + 2M(z - x) = 2Mx - M^2 + 2Mz - 2Mx \\ &= z^2 - z^2 + 2Mz - M^2 = z^2 - \underbrace{(z-M)^2}_{\geq 0} \end{aligned}$$

Case 2: $|x| \leq M$, but $|z| > M$

$$\begin{aligned} M(2z - M) &\overset{!}{\geq} x^2 + 2x(z-x) = x^2 + 2xz - 2x^2 = 2xz - x^2 \\ M(2z - M) &= 2zM - M^2 = \ldots = z^2 - (z-M)^2 \overset{!}{\geq} z^2 - (z-x)^2 \\ &\qquad\qquad\qquad\qquad (z-M)^2 \leq (z-x)^2 \end{aligned}$$

The last line is true because $z > M > x$.

(c) Download the provided measurements from the exercise homepage. Minimize the Huber norm using MATLAB. You do not need the Classification Toolbox. Use the MATLAB function `fminunc`.

See `linefitting.m`

(d) Compare the robust line fitting to a ordinary least-square approach. Find situations where the robust approach is superior. Show that due to convexity, the optimum is always found.

Exercise 2 A training set of N independent samples with feature vectors $\boldsymbol{a}_i \in \mathbb{R}^D$ and target variables $b_i \in \mathbb{R}$ is given. A linear model with the parameter $\boldsymbol{x} \in \mathbb{R}^D$ is assumed to estimate the target variable from the feature $b = \boldsymbol{x}^T \boldsymbol{a}$.

Ridge regression is least-squares linear regression with L_2-norm regularization. It is defined by the optimization problem

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmin}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \lambda \|\boldsymbol{x}\|_2^2 \quad , \tag{4}$$

with the design matrix $\boldsymbol{A} \in \mathbb{R}^{N \times D}$, $\boldsymbol{A}(i,j) = \boldsymbol{a}_i(j)$ and the target vector $\boldsymbol{b} \in \mathbb{R}^D$, $\boldsymbol{b}(i) = b_i$.

(a) Derive the solution of the ridge regression optimization problem.

Reformulate the optimization problem

$$\begin{aligned}
&\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2+\lambda\|\boldsymbol{x}\|_2^2\\
&=(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^T(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})+\lambda\boldsymbol{x}^T\boldsymbol{x}\\
&=\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}-2\boldsymbol{b}^T\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}^T\boldsymbol{b}+\lambda\boldsymbol{x}^T\boldsymbol{x}\\
&=\boldsymbol{x}^T\left(\boldsymbol{A}^T\boldsymbol{A}+\lambda\boldsymbol{I}\right)\boldsymbol{x}-2\boldsymbol{b}^T\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}^T\boldsymbol{b}
\end{aligned}$$

Compute the derivative w.r.t. $\boldsymbol{x}$

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{x}}\left(\boldsymbol{x}^T\left(\boldsymbol{A}^T\boldsymbol{A}+\lambda\boldsymbol{I}\right)\boldsymbol{x}-2\boldsymbol{b}^T\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}^T\boldsymbol{b}\right)&=0\\
2\left(\boldsymbol{A}^T\boldsymbol{A}+\lambda\boldsymbol{I}\right)\boldsymbol{x}-2\boldsymbol{A}^T\boldsymbol{b}&=0\\
\left(\boldsymbol{A}^T\boldsymbol{A}+\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{A}^T\boldsymbol{b}&=\boldsymbol{x}
\end{aligned}$$

(b) What is the effect of the regularization?

- Coefficients of $\boldsymbol{x}$ are forced to be smaller (shrinkage)
- The bias of the estimate is higher, but the variance is smaller
- More numerical stability (condition of matrix inversion is improved)

(c) Ridge regression can be motivated by Maximum A Posteriori (MAP) estimation. In MAP estimation, the a posteriori probability of the parameters after observing the training data is maximized $\boldsymbol{x}^*=\operatorname{argmax}_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{b})$. The assumption of Gaussian noise $p(b|\boldsymbol{x},\boldsymbol{a})=\mathcal{N}\left(b|\boldsymbol{x}^T\boldsymbol{a},\beta^{-1}\right)$ and a Gaussian prior for the parameters $p(\boldsymbol{x})=\mathcal{N}(\boldsymbol{x}|\boldsymbol{0},\alpha^{-1}\boldsymbol{I})$ is made. Show that MAP estimation in this setting is equivalent to ridge regression.

Define the a posteriori probability:

$$\begin{aligned}
&p(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{b})\propto p(\boldsymbol{b}|\boldsymbol{x},\boldsymbol{A})\cdot p(\boldsymbol{x})\\
&p(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{b})\propto\prod_{i=1}^{N}p(b_i|\boldsymbol{x},\boldsymbol{a}_i)\cdot p(\boldsymbol{x})\\
&p(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{b})\propto\prod_{i=1}^{N}\mathcal{N}\left(b_i|\boldsymbol{x}^T\boldsymbol{a}_i,\beta^{-1}\right)\cdot\mathcal{N}\left(\boldsymbol{x}|\boldsymbol{0},\alpha^{-1}\boldsymbol{I}\right)
\end{aligned}$$

Apply the logarithm and insert the formulas for the Gaussians:

$$\begin{aligned}
&\log p(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{b})\propto\sum_{i=1}^{N}\log\mathcal{N}\left(b_i|\boldsymbol{x}^T\boldsymbol{a}_i,\beta^{-1}\right)+\log\mathcal{N}\left(\boldsymbol{x}|\boldsymbol{0},\alpha^{-1}\boldsymbol{I}\right)\\
&\log p(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{b})\propto\sum_{i=1}^{N}\log\frac{1}{\sqrt{2\pi\beta^{-1}}}e^{-\frac{\beta}{2}\left(b_i-\boldsymbol{a}_i^T\boldsymbol{x}\right)^2}+\log\frac{1}{\sqrt{(2\pi)^D\alpha^{-D}}}e^{-\frac{\alpha}{2}\boldsymbol{x}^T\boldsymbol{x}}\\
&\log p(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{b})\propto-\frac{N}{2}\log 2\pi\beta^{-1}-\sum_{i=1}^{N}\frac{\beta}{2}\left(b_i-\boldsymbol{a}_i^T\boldsymbol{x}\right)^2-\frac{D}{2}\log 2\pi\alpha^{-1}-\frac{\alpha}{2}\boldsymbol{x}^T\boldsymbol{x}\\
&\log p(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{b})\propto-\sum_{i=1}^{N}\frac{\beta}{2}\left(b_i-\boldsymbol{a}_i^T\boldsymbol{x}\right)^2-\frac{\alpha}{2}\boldsymbol{x}^T\boldsymbol{x}+\text{const.}\\
&\log p(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{b})\propto-\frac{\beta}{2}(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^T(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})-\frac{\alpha}{2}\boldsymbol{x}^T\boldsymbol{x}+\text{const.}\\
&\log p(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{b})\propto-\frac{\beta}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2-\frac{\alpha}{2}\|\boldsymbol{x}\|_2+\text{const.}
\end{aligned}$$

Maximization of this probability is equivalent to minimization of Eq.4.