Lehrstuhl für STEUERUNGS- UND REGELUNGSTECHNIK Technische Universität München Prof. Dongheui Lee	**MACHINE LEARNING IN ROBOTICS** Exercises Non-parametric Density Est & Dim Reduction

Exercise 1

In density estimation with Parzen windows, we distribute hypercubes of fixed size in the dataspace and the number of samples involved in these hypercubes is estimated based on the given data distribution. Assume that an 1D dataset is observed and the number of samples per hypercube is given by $K = \sum_{i=1}^{n} k(\frac{x-x^{(i)}}{h})$ where

$$k(u) = \begin{cases} 1 & |u| \leq 1/2 \\ 0 & otherwise \end{cases}$$

a) Prove that the pdf of the data distribution is given by $p(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \cdot k(\frac{x-x^{(i)}}{h})$.

b) Analyze the effect that the window width has on the estimation of the function $p(x)$.

Solution

(a)The general formula giving the pdf of the data distribution is given by:

$$p(x) = \frac{K}{nV} = \frac{\sum_{i=1}^{n} k(\frac{x-x^{(i)}}{h})}{nh}$$

(b)

$$p_n(x) = \frac{1}{nh} \sum_{i=1}^{n} k(\frac{x-x^{(i)}}{h})$$

h affects the amplitude of the density function and its width.

- if h large, the amplitude decreases and the density function $p(x)$ becomes smoother and maybe oversmoothed.
- if h small, $p(x)$ becomes high and this increase occurs where x is very near to the $x^{(i)}$. Thus, $p(x)$ approaches a Dirac delta function centered at $x^{(i)}$ and $p(x)$ arises from a superposition of sharp pulses centered too close to the sample points.

Exercise 2

Consider a histogram-like density model in which the space x is divided into fixed regions for which the density $p(x)$ takes the constant value h_k over the k^{th} region, and that the volume of region k is denoted Δ_k . Suppose we have a set of n observations of x such that n_k of these observations fall in region k. Using a Lagrange multiplier to enforce the normalization constraint on the density, derive an expression for the maximum likelihood estimator for the h_k.

Solution

The value of the density $p(x)$ at a point $x^{(i)}$ is given by $h_{j(i)}$, where the notation $j(i)$ denotes that data point $x^{(i)}$ falls within region j. Thus the log-likelihood function takes the form

$$l = \sum_{i=1}^{n} \ln p(x^{(i)}) = \sum_{i=1}^{n} \ln h_{j(i)}$$

We now need to take account of the constraint that $p(x)$ must integrate to unity. Since $p(x)$ has the constant value h_r over region r, which has volume Δ_r , the normalization constraint becomes $\sum_r h_r \Delta_r = 1$. Introducing a Lagrange multiplier λ we then maximize the function

$$\sum_{i=1}^{n} \ln h_{j(i)} + \lambda \left(\sum_r h_r \Delta_r - 1 \right)$$

Taking partial derivative with respect to h_k and evaluating it to zero gives:

$$0 = \frac{n_k}{h_k} + \lambda \Delta_k$$

$$0 = n_k + \lambda \Delta_k h_k$$

where n_k denotes the total number of data points falling within region k.

Now making use of the normalization constraint,

$$0 = \sum_r n_r + \lambda (\sum_r h_r \Delta_r)$$

$$\Rightarrow \lambda = -n$$

Eliminating λ then gives our final result for the maximum likelihood solution for h_k in the form

$$h_k = \frac{n_k}{n} \frac{1}{\Delta_k}$$

Note that, for equal sized bins $\Delta_k = \Delta$ we obtain a bin height h_k which is proportional to the fraction of points falling within that bin, as expected.

Exercise 3

Show that the K-nearest-neighbour density model defines an improper distribution whose integral over all space is divergent.

Solution

Lets consider a 1-dimensional problem with fixed K. The density estimate is given by:

$$p(x) = \frac{K}{nV}$$

where n is the total number of data points and V is the width of the bin. Now if we place the bin at the right most point x_r, then with width Δ_r it encloses k points. Now if we move by an amount δ to the right of x_r, the bin width will increase by 2δ and the density estimate of that point will be:

$$p(x_r + \delta) = \frac{K}{n(\Delta_r + 2\delta)}$$

To be a proper density estimate, a probability density function should integrate to 1. Now for testing the convergence, if we integrate δ from $0 \rightarrow \infty$ we get

$\int\limits_{\delta=0}^{\infty} p(x_r+\delta)d\delta = \int\limits_{\delta=0}^{\infty} \frac{K}{n(\Delta_r+2\delta)}d\delta = \frac{K}{2n}\ln|n(\Delta_r+2\delta)|\Big|_{\delta=0}^{\infty} = \infty$

$\Bigg($ Integral of the above form is calculated as: $\int \frac{c}{ax+b}dx = \frac{c}{a}\ln|ax+b| + C$ $\Bigg)$

and thus the integral is divergent and the density model is an improper distribution.

Exercise 4

Suppose that for an image recognition task, we want to extract the principal components from a set of n images on a mobile robot. We reshaped each image into a column vector and now we have $d \times n$-dimensional data X which has n samples of $d \times 1$-dimensional vectors. These vectors correspond to image data and the dimension $d \gg n$. The principal components correpond to eigenvector of covariance matrix of C

$$C = \frac{1}{n-1} X_c {X_c}^\top \tag{1}$$

where X_c is obtained by subtracting the mean vector from X. Since for a large value of d (for an image of size 640×480, $d = 307200$), this computation can easily hang the On-board computer.

A useful trick is to calculate the eigenvector of C_1 which is $n \times n$

$$C_1 = \frac{1}{n-1} {X_c}^\top X_c \tag{2}$$

and now if v_1 is an eigenvector of C_1 with corresponding eigenvalue λ then $v = X_c v_1$ is an eigenvector of C with corresponding eigenvalue λ. Show that this claim is true.

Solution

If v_1 is an eigenvector of C_1 then

$C_1 v_1 = \lambda v_1 \Rightarrow \lambda v_1 = \frac{1}{n-1} {X_c}^\top X_c v_1$

Multiplying bothsides by X_c we get

$\lambda X_c v_1 = \frac{1}{n-1} X_c {X_c}^\top X_c v_1$

$\lambda v = \frac{1}{n-1} X_c {X_c}^\top v$

$\lambda v = C v$

Exercise 5

Use prove by induction to show that the linear projection onto an M-dimensional subspace that maximizes the variance of the projected data is defined by the M eigenvectors of the data covariance matrix S, corresponding to the M largest eigenvalues.

Solution

Formula for covariance matrix is $\Sigma = E\left[(X - E[X])(X - E[X])^{\top}\right] = \frac{\sum_{i=1}^{n}(X^{(i)}-E[X])(X^{(i)}-E[X])^{\top}}{n-1}$

For induction we have first to proof for $M = 1$ (dimensionality reduction to 1 dimension), then assuming that for M (dimensionality reduction to M dimensions) the condition holds, we prove it for $M + 1$.
For $M = 1$:
Our projection on a given vector e_1 is defined by $y = e_1^{\top} x^{(n)}$. Then our projected mean and covariance would be:

$$\begin{aligned} Projected\ mean\ &= \frac{\sum e_1^{\top} x^{(i)}}{n} = e_1^{\top}\overline{x} \qquad (3) \\ Projected\ variance\ &= \frac{\sum \{e_1^{\top} x^{(i)} - e_1^{\top}\overline{x}\}^2}{n-1} = \frac{\sum \{e_1^{\top}(x^{(i)} - \overline{x})\}^2}{n-1} \\ &= \frac{\sum \left[e_1^{\top}(x^{(i)} - \overline{x})\right]\left[e_1^{\top}(x^{(i)} - \overline{x})\right]^{\top}}{n-1} \\ &= \frac{\sum e_1^{\top}(x^{(i)} - \overline{x})(x^{(i)} - \overline{x})^{\top} e_1}{n-1} \\ &= e_1^{\top}\frac{\sum \left[(x^{(i)} - \overline{x})(x^{(i)} - \overline{x})^{\top}\right]}{n-1} e_1 = e_1^{\top} S e_1 \qquad (4) \end{aligned}$$

S is the covariance matrix and we need to maximize the variance. Therefore, adding a Lagrange multiplier to constraint our vectors norm, we can maximize the variance with:

$$\begin{aligned} u &= e_1^{\top} S e_1 - \lambda_1(e_1^{\top} e_1 - 1) \qquad (5) \\ \frac{\partial u}{\partial e_1} &= 2Se_1 - 2\lambda_1 e_1 = 0 \qquad (6) \\ Se_1 &= \lambda_1 e_1 \qquad (7) \end{aligned}$$

The vector that maximizes this expression ($e_1^{\top} S e_1 = \lambda e_1^{\top} e_1$) is the eigenvector of S that has the highest eigenvalue, and is called the first principal component.

Now we suppose that this condition holds for M and we try to prove the same for $M + 1$.
The variance in the $M + 1$ direction is given by:

$$Projected\ variance\ = e_{M+1}^{\top} S e_{M+1} \qquad (8)$$

To maximize it, we just need to add the same constraint as before, and an orthogonality constraint for each one of the previous M orthogonal components. Therefore, we also add M Lagrange multipliers $\eta_1..\eta_M$ for each orthogonal vector:

$$\begin{aligned} u &= e_{M+1}^{\top} S e_{M+1} - \lambda_{M+1}(e_{M+1}^{\top} e_{M+1} - 1) + \sum_{o=1}^{M} \eta_o e_{M+1}^{\top} e_o \qquad (9) \\ \frac{\partial u}{\partial e_{M+1}} &= 2Se_{M+1} - 2\lambda_{M+1} e_{M+1} + \sum_{o=1}^{M} \eta_o e_o = 0 \qquad (10) \end{aligned}$$

If we now left-multiply this result with one of the previous $e_j^\top$, we get:

$$e_j^\top 2Se_{M+1} - e_j^\top 2\lambda_{M+1}e_{M+1} + \sum_{o=1}^{M} e_j^\top \eta_o e_o = 0 \tag{11}$$

$e_j^\top 2Se_{M+1} = (e_j^\top 2Se_{M+1})^\top = 2e_{M+1}^\top S^\top e_j = 2e_{M+1}^\top Se_j = 2\lambda_j e_{M+1}^\top e_j = 0$

But as our new vector e_{M+1} must be orthogonal to any of the previous one, which are already orthogonal between themselves, we get:

$$e_j^\top \eta_j e_j = 0 \tag{12}$$

Applying this for every previous orthogonal vector, we realize that $\eta_j = 0$ for every $j = 1..M$. Knowing that our maximization equation is:

$$Se_{M+1} = \lambda_{M+1}e_{M+1} \tag{13}$$

And to maximize this equation e_{M+1} must be the eigenvector with the highest eigenvalue, different from the previous M.