IFT6390-fundamentals-of-machine-learning
Assignment 1

Jonathan Guymont

September 26, 2018

Parametric Gaussian density estimation, v.s. Parzen window density estimation

Question 1

(a) The density estimation of an isotropic gaussian

$$N_{\mu,\sigma}(x) = \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d}\exp(-\frac{1}{2}\frac{||x-\mu||^2}{\sigma^2})$$

and the parameters are the mean μ and the covariance matrix Σ. The dimension of μ is d and the dimension of Σ is $d \times d$.

(b) The mean μ is given by $(\mu_1, ..., \mu_d)^\top$ where

$$\mu_i = \frac{1}{n}\sum_{j=1}^{n} x_i^{(j)}, i = 1, ..., d.$$

Since the gaussian is isotropic, we have that $\Sigma_{i,j} = 0$ for all $i \neq j$ and $\Sigma_{i,i} = \Sigma_{j,j}$ for all i, j (e.g. each feature of x has the same variance). The log-likelihood estimator of the variance in the isotropic case is given by $\hat{\sigma}_{MLE}I$, where $\hat{\sigma}_{MLE}$ is the average over the empirical variances of the features

$$\sigma = \frac{1}{d}\sum_{i=1}^{d}\sigma_{i,i}$$

where

$$\sigma_{i,i} = \frac{1}{n}\sum_{j=1}^{n}(x_i^{(j)} - \mu_i)^2$$

(c) The complexity of computing the average over a feature $O(n)$. The complexity of computing the variance over a feature is $O(n)$. Both of this computation needs to be done d

times. Thus, the complexity of the method is $O(dn)$

(d) The density estimation of an isotropic gaussian

$$\hat{p}(x) = \frac{1}{(2\pi)^{\frac{d}{2}}\hat{\sigma}^d} \exp(-\frac{1}{2}\frac{||x-\hat{\mu}||^2}{\hat{\sigma}^2})$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the estimators in (c).

(e) The complexity is $O(d)$ since computing the dot product between x and μ needs d substraction, d mutiplication (square), and d addition, all linear in d.

Question 2

(a) The algorithm learn (memorize) the dataset.

(b) The Gaussian Parzen window density estimation of x is given by

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} K(x^{(i)}, x)$$

where

$$K(x^{(i)}, x) = \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d} \exp\left(-\frac{1}{2}\frac{||x^{(i)}-x||^2}{\sigma^2}\right)$$

If we replace the kernel in the density we obtain

$$\hat{p}(x) = \frac{1}{n(2\pi)^{\frac{d}{2}}\sigma^d} \sum_{i=1}^{n} \exp\left(-\frac{1}{2}\frac{||x^{(i)}-x||^2}{\sigma^2}\right)$$

(c) The complexity is $O(dn)$ since the complexity of computing the euclidian distance is $O(d)$ and we sum over n examples.

Question 3

(a) The Parzen as more capacity because its memorizing all the dataset in order to make a prediction. The parametric gaussian is compressing all the information about the underlying distribution of the data in two variable si it needs to summarize the important information in far less parameters.

(b) If the variance of the Parzen method is low, it is likely to overfit.

(c) The parameters of the parametric Gaussian are learned parameters. We don't have to use optimization technic like the one use in deep learning since it is possible to obtain an anlytic (or closed) solution for the parameters that minimized the log-likelihood in the case of the parametric Gaussian density estimation. The variance in KDE is not selected by optimization of a learning criteria.

Question 4

(a) The density of a diagonal multivariate Gaussian is given by

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{1}$$

where $\boldsymbol{x} = (x_1, ..., x_d)^\top$ and $\boldsymbol{\mu} = (\mu_1, ..., \mu_d)^\top$ and $\Sigma = (\sigma_{ij})_{i=1,..,d;j=1,...,d}$ are the parameters. The dimension of μ is d and the dimentsion of Σ is $d \times d$.

(b) In this question, σ_{ij} is the covariance between X_i and X_j as oppose to the standard deviation oftenly denote by σ. A set of random variables $X_1, ..., X_d$ is independent if the joint density is equal to the product of the marginal density, i.e. $p(X_1, ..., X_d) = p(X_1) \cdots p(X_d)$. Also, the marginal density is given by

$$p(X_i) = (2\pi\sigma_i)^{-\frac{1}{2}} \exp\left(\frac{(X_i-\mu_i)^2}{2\sigma_{ii}}\right), i = 1, ..., d$$

Since Σ is diagonal, its inverse is given by

$$\Sigma^{-1} = \begin{pmatrix} \sigma_{11}^{-1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{dd}^{-1} \end{pmatrix}$$

where σ_{ii} is the variance of X_i (not the standard deviation!). Also, the numerator in 1 can be writen has

$$\begin{aligned}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) &= \begin{bmatrix}(x_1-\mu_1)\sigma_{11}^{-1} \cdots (x_d-\mu_d)\sigma_{dd}^{-1}\end{bmatrix} \begin{bmatrix}(x_1-\mu_1) \\ \vdots \\ (x_d-\mu_d)\end{bmatrix} \\ &= \sum_{i=1}^{d}(x_i-\mu_i)^2\sigma_{ii}^{-1}\end{aligned} \tag{2}$$

Also, the determinant of a diagonal matrix is given by the product of the element of the diagonal. Hence, we can rewrite the joint density has

$$\begin{aligned}p(\boldsymbol{x}) =& (2\pi)^{-\frac{d}{2}}\sqrt{\Pi_{i=1}^{d}\sigma_{ii}^{-1}} \exp\left(-\frac{1}{2}\sum_{i=1}^{d}(x_i-\mu_i)^2\sigma_{ii}^{-1}\right) \\ =& \Pi_{i=1}^{d}(2\pi)^{-\frac{1}{2}}\sigma_{ii}^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x_i-\mu_i)^2\sigma_{ii}^{-1}\right)\end{aligned} \tag{3}$$

which is the product of the marginal density.
(c) We want to minimize

$$L = -\frac{1}{n}\sum_{i=1}^{n}\log p(x^{(i)}) \tag{4}$$

with

$$\log p(x) = -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu) \tag{5}$$

Since the first term of 5 is constant with respect to the learn parameters, the function to minimize is

$$L = \frac{1}{2}\log|\Sigma| + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu) + const \tag{6}$$

(d) We first optimize in μ.

$$\begin{aligned}
\frac{\partial}{\partial\mu}L =& \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left(-\Sigma^{-1}(x-\mu) - (x-\mu)^\top\Sigma^{-1}\right)\\
=& \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left(-\Sigma^{-1}(x-\mu) - \Sigma^{-1}(x-\mu)\right)\\
=& \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left(-2\Sigma^{-1}(x-\mu)\right)\\
=& -\frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}(x-\mu)
\end{aligned} \tag{7}$$

Setting the derivative equal to 0, we have

$$\begin{aligned}
-\Sigma^{-1}\frac{1}{n}\sum_{i=1}^{n}(x-\mu) =& 0\\
\Sigma\cdot\Sigma^{-1}\frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}(x-\mu) =& \Sigma\cdot 0\\
\frac{1}{n}\sum_{i=1}^{n}x =& \frac{1}{n}\sum_{i=1}^{n}\mu\\
\frac{1}{n}\sum_{i=1}^{n}x =& \mu
\end{aligned} \tag{8}$$

We now compute the derivative with respect to Σ. Note that $(x-\mu)^\top\Sigma^{-1}(x-\mu)$ is a scalar, hence it is equal to its trace. Also, using the fact that $\mathrm{Tr}AB = \mathrm{Tr}BA$, we have that $(x-\mu)^\top\Sigma^{-1}(x-\mu) = \mathrm{Tr}(x-\mu)^\top\Sigma^{-1}(x-\mu) = \mathrm{Tr}(x-\mu)(x-\mu)^\top\Sigma^{-1} = \mathrm{Tr}\Sigma^{-1}(x-\mu)(x-\mu)^\top$. Replacing in equation 6 we have

$$\begin{aligned}
L =& \frac{1}{2}\log|\Sigma| + \frac{1}{2n}\sum_{i=1}^{n}(x-\mu)^\top\Sigma^{-1}(x-\mu) + const\\
=& \frac{1}{2}\log|\Sigma| + \frac{1}{2n}\sum_{i=1}^{n}\mathrm{Tr}\Sigma^{-1}(x-\mu)(x-\mu)^\top + const\\
=& \frac{1}{2}\log|\Sigma| + \frac{1}{2n}\mathrm{Tr}\Sigma^{-1}\sum_{i=1}^{n}(x-\mu)(x-\mu)^\top + const
\end{aligned} \tag{9}$$

To simplify the notation, let $S = \sum_{i=1}^{n}(x-\mu)(x-\mu)^\top$. We now have

$$L = \frac{1}{2}\log|\Sigma| + \frac{1}{2n}\mathrm{Tr}\Sigma^{-1}S + const \tag{10}$$

Setting $\Lambda = \Sigma^{-1}$ we have [1]

$$\begin{aligned} L =& \frac{1}{2}\log|\Lambda^{-1}| + \frac{1}{2n}\mathrm{Tr}\Lambda S + const \\ =& -\frac{1}{2}\log|\Lambda| + \frac{1}{2n}\mathrm{Tr}\Lambda S + const \end{aligned} \tag{11}$$

Taking the derivative w.r.t Σ, setting it to zero, and using the fact that the derivative of the log of the determinant a matrix is the inverse of the said matrix, and that $\nabla_B \mathrm{Tr}AB = A$ we have

$$\begin{aligned} -\Lambda^{-1} + \frac{1}{n}S =& 0 \\ \Lambda^{-1} =& \frac{1}{n}S \\ \Sigma =& \frac{1}{n}S \\ \Sigma =& \frac{1}{n}\sum_{i=1}^{n}(x-\mu)(x-\mu)^\top \end{aligned} \tag{12}$$

[1] This trick comes from University of British Columbia course notes: https://www.cs.ubc.ca/ schmidtm/Courses/540-W17/L9.pdf