

# IFT6390-fundamentals-of-machine-learning

## Assignment 1

Jonathan Guymont

September 23, 2018

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

#### Question 1

(a) The density estimation of an isotropic gaussian

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

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

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

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

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  $O(n)$

(d) The density estimation of an isotropic gaussian

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

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

(e) The complexity is  $O(1)$ .

## Question 2

(a) There is no learning since the prediction is only a function of  $\sigma$ .

(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

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

(c) The complexity is  $O(n)$ .

## Question 3

(a) The parametric Gaussian has more capacity because it has more parameters. Also, parameters of the parametric Gaussian are learned based on the training set, while the parameter of the KDE is more like a hyperparameter making it more difficult to overfit the training set with the KDE.

(b) If you don't have enough data, the parametric Gaussian is likely to overfit.

(c) The parameters of the parametric Gaussian are learned parameters. We don't have to use optimization technique like the one used in deep learning since it is possible to obtain an analytic (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

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

where  $\mu$  and  $\Sigma$  are the parameters. The dimension of  $\mu$  is  $d$  and the dimension of  $\Sigma$  is  $d \times d$ .

(b) Two variable  $X$  and  $Y$  are independent if  $p(X, Y) = p(X)p(Y)$ .

(c) We want to minimize

$$L = -\frac{1}{n} \sum_{i=1}^n \log p(x^{(i)}) \quad (2)$$

with

$$\log p(x) = -\frac{d}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \quad (3)$$

Since the first term of 3 is constant with respect to the learn parameters, the function to minimize is

$$L = \frac{1}{2} \log |\Sigma| + \frac{1}{n} \sum_{i=1}^n \frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) + \text{const} \quad (4)$$

(d) We first optimize in  $\mu$ .

$$\begin{aligned} \frac{\partial}{\partial \mu} L &= \frac{1}{n} \sum_{i=1}^n \frac{1}{2} (-\Sigma^{-1} (x - \mu) - (x - \mu)^\top \Sigma^{-1}) \\ &= \frac{1}{n} \sum_{i=1}^n \frac{1}{2} (-\Sigma^{-1} (x - \mu) - \Sigma^{-1} (x - \mu)) \\ &= \frac{1}{n} \sum_{i=1}^n \frac{1}{2} (-2\Sigma^{-1} (x - \mu)) \\ &= -\frac{1}{n} \sum_{i=1}^n \Sigma^{-1} (x - \mu) \end{aligned} \quad (5)$$

Setting the derivative equal to 0, we have

$$\begin{aligned} -\Sigma^{-1} \frac{1}{n} \sum_{i=1}^n (x - \mu) &= 0 \\ \Sigma \cdot \Sigma^{-1} \frac{1}{n} \sum_{i=1}^n \Sigma^{-1} (x - \mu) &= \Sigma \cdot 0 \\ \frac{1}{n} \sum_{i=1}^n x &= \frac{1}{n} \sum_{i=1}^n \mu \\ \frac{1}{n} \sum_{i=1}^n x &= \mu \end{aligned} \quad (6)$$

We now compute the derivative with respect to  $\Sigma$ . Note that  $(x - \mu)^\top \Sigma^{-1} (x - \mu)$  is a scalar, hence it is equal to its trace. Also, using the fact that  $\text{Tr} AB = \text{Tr} BA$ , we have that  $(x - \mu)^\top \Sigma^{-1} (x - \mu) = \text{Tr}(x - \mu)^\top \Sigma^{-1} (x - \mu) = \text{Tr}(x - \mu)(x - \mu)^\top \Sigma^{-1} = \text{Tr} \Sigma^{-1} (x - \mu)(x - \mu)^\top$ . Replacing in equation 4 we have

$$\begin{aligned} L &= \frac{1}{2} \log |\Sigma| + \frac{1}{2n} \sum_{i=1}^n (x - \mu)^\top \Sigma^{-1} (x - \mu) + \text{const} \\ &= \frac{1}{2} \log |\Sigma| + \frac{1}{2n} \sum_{i=1}^n \text{Tr} \Sigma^{-1} (x - \mu)(x - \mu)^\top + \text{const} \\ &= \frac{1}{2} \log |\Sigma| + \frac{1}{2n} \text{Tr} \Sigma^{-1} \sum_{i=1}^n (x - \mu)(x - \mu)^\top + \text{const} \end{aligned} \quad (7)$$

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

$$L = \frac{1}{2} \log |\Sigma| + \frac{1}{2n} \text{Tr} \Sigma^{-1} S + \text{const} \quad (8)$$

Setting  $\Lambda = \Sigma^{-1}$  we have <sup>1</sup>

$$\begin{aligned} L &= \frac{1}{2} \log |\Lambda^{-1}| + \frac{1}{2n} \text{Tr} \Lambda S + \text{const} \\ &= -\frac{1}{2} \log |\Lambda| + \frac{1}{2n} \text{Tr} \Lambda S + \text{const} \end{aligned} \quad (9)$$

Taking the derivative w.r.t  $\Sigma$ , 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  $\nabla_B \text{Tr} AB = A$  we have

$$\begin{aligned} -\Lambda^{-1} + \frac{1}{n} S &= 0 \\ \Lambda^{-1} &= \frac{1}{n} S \\ \Sigma &= \frac{1}{n} S \\ \Sigma &= \frac{1}{n} \sum_{i=1}^n (x - \mu)(x - \mu)^\top \end{aligned} \quad (10)$$

---

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