IFT 6390 Homework 1

Author

September 20, 2018

Problem 1. Small exercise on probabilities

Proof. p □

Problem 3.1 (a). Name parameters

Proof. the parameters are σ and μ
$\mu \in (d, 1)$
The matrix Σ is d by d, where the diagonal terms are σ □

Problem 3.1 (b). Equation for optimal parameters

Proof.

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\sigma = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T$$

□

Problem 3.1 (c). What is the algorithmic complexity of this training method, i.e. of the method calculating these parameters?

Proof. To computer μ:
for i = 1 to n
sum each x_i with d components
endfor
Total is $O(nd)$
To computer σ:
for i = 1 to n
subtract μ from each x_i with d components
repeat the last operation with transpose multiply d by 1 and 1 by d vectors, result in a d by d matrix, which is d^2 operations endfor
Total is $O(nd^2 + d + d) = O(nd^2)$

□

Problem 3.2(a). Suppose that the user has fixed σ What does the "training/learning" phase of these Parzen windows consist of?

Proof. Center at each x, calculate normal density function with σ □

Problem 3.2(b). For a test point x, write in a single detailed formula (i.e. with exponentials), the function that will give the probability density predicted at point x: $p_{parzen}(x) =$?

Proof.

$$\hat{p}_{Parzen}(x) = \sum_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{d/2}} exp(\frac{1}{2}\frac{|x-\mu|^2}{\sigma})$$

□

Problem 3.2(c). What is the algorithmic complexity for calculating this prediction at each new point x?

Proof. $O(N)$, since there are n terms and all have fixed σ and μ □

Problem 3.3(a). Which one of these two approaches (parametric Gaussian v.s. Parzen Gaussian kernel) has the highest capacity (in other words, higher expressivity)? Explain

Proof. Parzen □

Problem 3.3(b). With which one of these approaches, and in which scenario, are we likely to be over-fitting (i.e. memorizing the noise in our data)?

Proof. Parzen □

Problem 3.3(c). Hyperparam vs param

Proof. to do □

Problem 3.4 (a). Express the equation of a diagonal Gaussian density in Rd. Specify what are its parameters and their dimensions.

Proof.

$$\prod_{i=1}^{D} (2\pi\sigma_i)^{\frac{1}{2}} exp\{-\sum_{i=1}^{D} \frac{1}{2\sigma_i^2}(x_i - \mu_i)^2\}$$

□

Problem 3.4 (b). Show that the components of a random vector following a diagonal Gaussian distribution are independent random variables.

Proof. First, let us walk through the idea. We would like to prove the independence by proving that the product of each normal distribution's density function made up of the each component of the random vector is equal to the diagonal Gaussian distribution's density function, namely the random vector's density distribution.
The Gaussian density function for dataset D is defined as

$$p(\boldsymbol{x}) = \mathcal{N}_{\mu,\Sigma}(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{det(\Sigma)}} exp(-\frac{1}{2}(\boldsymbol{x}-\mu)^T\Sigma^{-1}(\boldsymbol{x}-\mu))$$

Since we have a diagonal Σ, it can be written in the form of a summation of the diagonal terms in the power, and then transformed into a product of exponents. Note that σ_{ii} here refers to the diagonal terms.

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{det(\Sigma)}} exp(\sum_{i=1}^{d} -\frac{1}{2\sigma_{ii}^2}(x_i-\mu)^2)$$

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{det(\Sigma)}} \prod_{i=1}^{d} exp - (\frac{1}{2\sigma_{ii}^2}(x_i-\mu)^2)$$

we can also express the constant term in a product form

$$\frac{1}{(2\pi)^{d/2}\sqrt{det(\Sigma)}} = \prod_{i=1}^{d} \frac{1}{(2\pi)^{1/2}\sigma_{ii}} = \prod_{i=1}^{d} \frac{1}{(2\pi\sigma_{ii}^2)^{1/2}}$$

also, note that the determinant of a diagonal matrix is the product of the diagonal terms

$$det(\Sigma) = \prod_{i=1}^{d} \sigma_{ii}^2$$

Now, we have the expression for $p(\boldsymbol{x})$

$$p(\boldsymbol{x}) = \prod_{i=1}^{d} \frac{1}{(2\pi\sigma_{ii}^2)^{1/2}} \prod_{i=1}^{d} exp(-\frac{1}{2\sigma_{ii}^2}(x_i-\mu)^2) = \prod_{i=1}^{d} \frac{1}{(2\pi\sigma_{ii}^2)^{1/2}} exp(-\frac{1}{2\sigma_{ii}^2}(x_i-\mu)^2)$$

Note that $\frac{1}{(2\pi\sigma_{ii}^2)^{1/2}} exp(\frac{1}{2\sigma_{ii}^2}(x_i-\mu)^2)$ is a Gaussian of $\mathcal{N}(\sigma_{ii},\mu)$
Therefore, we have

$$p(\boldsymbol{x}) = \prod_{i=1}^{d} \mathcal{N}(\sigma_{ii},\mu)$$

□

Problem 3.4(c). Using ? log p(x) as the loss, write down the equation correspond- ing to the empirical risk minimization on the training set D

Proof. From part b, we have

$$p(\boldsymbol{x}) = \prod_{i=1}^{d} \frac{1}{(2\pi\sigma_{ii}^2)^{1/2}} exp(\frac{1}{2\sigma_{ii}^2}(x_i - \mu)^2)$$

$$-logp(X) = -log(\prod_{i=1}^{d} \frac{1}{(2\pi\sigma_{ii}^2)^{1/2}} exp(-\frac{1}{2\sigma_{ii}^2}(x_i - \mu)^2))$$

$$-logp(X) = \sum_{i=1}^{d} -log(\frac{1}{(2\pi\sigma_{ii}^2)^{1/2}} exp(-\frac{1}{2\sigma_{ii}^2}(x_i - \mu)^2))$$

$$-logp(X) = \sum_{i=1}^{d} -log(\frac{1}{(2\pi\sigma_{ii}^2)^{1/2}} - \sum_{i=1}^{d} log(exp(-(\frac{1}{2\sigma_{ii}^2}(x_i - \mu)^2)))$$

$$-logp(X) = \sum_{i=1}^{d} log((2\pi\sigma_{ii}^2)^{1/2} + \sum_{i=1}^{d} (\frac{1}{2\sigma_{ii}^2}(x_i - \mu)^2))$$

$$\hat{R}(f, D) = \frac{1}{|D|} \sum_{(x,y)\in D} L(f(x), y)$$

$$\hat{R}(f, D) = \frac{1}{|D|} \sum_{(x,y)\in D} -log(p(X))$$

$$\hat{R}(f, D) = \frac{1}{|D|} \sum_{(x,y)\in D} \sum_{i=1}^{d} log((2\pi\sigma_{ii}^2)^{1/2} + \sum_{i=1}^{d} (\frac{1}{2\sigma_{ii}^2}(x_i - \mu)^2))$$

□