IFT 6390 Homework 1

Arlie Coles (20121051) and Yue (Violet) Guo (20120727)

September 26, 2018

Problem 1. *Small exercise on probabilities.*

Solution. First, we let:

- the probability of a woman having cancer be $P(C) = 0.015$,
- the probability of a test for cancer being positive, given that a woman has cancer, be $P(T|C) = 0.87$,
- the probability of a test for cancer being positive, given that a woman does not have cancer, be $P(T|\neg C) = 0.096$.

Then, to find the probability that a woman who has received a positive result actually has cancer, $P(C|T)$, we apply Bayes' Rule:

$$\begin{aligned} P(C|T) &= \frac{P(T|C)P(C)}{P(T)} \\ &= \frac{P(T|C)P(C)}{P(T|\neg C)P(\neg C) + P(T|C) + P(C)} \\ &= \frac{(0.87)(0.015)}{(0.87 \times 0.015 + 0.096 \times (1 - 0.015))} \\ &= 0.01213 \end{aligned}$$

Therefore, the doctors surveyed ought to have responded with *F) Less than 10%*.

Problem 2. *Curse of dimensionality and geometric intuition in higher dimensions.*

Solution.

1. Volume can be generalized in d dimensions as $V = c^d$.

2. Since the probability density is zero anywhere outside the cube, we know that the integral of the probability density function over the volume must be equal to 1:

$$1 = \int_V p(x)dx$$

   And since the probability distribution is uniform, we can move it outside of the integral as it is a constant:

$$\begin{aligned} 1 &= p(x) \int_V dx \\ &= p(x)V \\ &= p(x)c^d \\ \Rightarrow p(x) &= \frac{1}{c^d}. \end{aligned}$$

3. We can define the volume of the interior cube as $(0.94c)^d$. Since the volume of the exterior cube is c^d, we can define the volume of the outer shell as $c^d - (0.94c)^d = c^d(1 - 0.94^d)$.

   Then, the probability of a generated point x falling within the shell is:

$$\begin{aligned} P(\text{shell}) &= \int_V p(x)dx \\ &= \frac{1}{c^d} \int_V dx \\ &= \frac{1}{c^d} c^d(1 - 0.94^d) \\ &= 1 - 0.94^d. \end{aligned}$$

   The probability of a generated point x falling in the smaller interior hypercube is:

$$
\begin{aligned}
P(\text{interior}) &= \int_V p(x)dx \\
&= \frac{1}{c^d}\int_V dx \\
&= \frac{1}{c^d}(0.94^d)(c^d) \\
&= 0.94^d.
\end{aligned}
$$

4. Changing values of d:

$$
\begin{array}{lll}
d = 1: & 1 - 0.94^1 & = 0.06 \\
d = 2: & 1 - 0.94^2 & = 0.1164 \\
d = 3: & 1 - 0.94^3 & = 0.1694 \\
d = 5: & 1 - 0.94^5 & = 0.2661 \\
d = 10: & 1 - 0.94^{10} & = 0.4614 \\
d = 100: & 1 - 0.94^{100} & = 0.9979 \\
d = 1000: & 1 - 0.94^{1000} & \approx 1
\end{array}
$$

5. In higher dimensions, it seems that the distribution of points is much more concentrated around the "edges" of the given "space", which runs counter to intuition in smaller dimensions about uniform distributions, where we suppose that there is not a relationship between location in the space and the value given by the probability density function.

Problem 3. *Parametric Gaussian density estimation vs. Parzen window density estimation*

Solution.

1. (a) An isotropic Gaussian has two parameters:
   - μ, the mean, of dimension d, and
   - σ^2, the variance, of dimension 1. (Note that the covariance matrix Σ is of dimension $d \times d$, where the diagonal terms are all equal to one σ^2 value.)

   (b) For calculating the mean:

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

   For calculating the variance:

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

   Then, we arrive at σ^2 by decomposing the resulting matrix (which is all zeros excepting the σ^2's on the diagonal, since the Gaussian is isotropic) into the following form, where the value of σ^2 is apparent:

$$\Sigma = \sigma^2 I$$

   (c) We can describe the complexity-affecting parts of the μ-calculating algorithm like so:

```
for i = 1 to n:                                          \\ O(n)
  add x_i to the running total, which has d components   \\ O(d)
```

   This is $O(nd)$.

   We can describe the complexity-affecting parts of the σ-calculating algorithm like so:

```
for i = 1 to n:                                          \\ O(n)
  subtract mu from x_i, both with d components           \\ O(d)
   repeat the previous operation                         \\ O(d)
   take its transpose                                    \\ O(1)
   multiply the resulting dx1 and 1xd vectors to get a dxd matrix
       \\ O(d^2)
```

This is $O(n(d+d+1+d^2)) = O(n(d+d^2)) \approx O(nd^2)$.

The complexity of calculating both parameters is then $O(nd + nd + nd^2) = O(nd + nd^2) = O(n(d+d^2)) \approx O(nd^2)$.

(d) The probability density function is:

$$\begin{aligned}\hat{p}_{gauss-isotrop}(x) &= \mathcal{N}_{\mu,\sigma^2} \\ &= \frac{1}{(2\pi)^{d/2}\sigma^d} \exp\left(\frac{-1}{2}\frac{||x-\mu||^2}{\sigma^2}\right)\end{aligned}$$

(e) Calculating a prediction from $p(x)$ at a given x only uses one data point, so it is not dependent on n. Subtracting μ from x in the exponent, however, is dependent on d since each has d components. Therefore the complexity is $O(d)$.

2. (a) The "training/learning" phase for the Parzen method consists of loading each training point. We center each kernel Gaussian on each training point, which becomes the μ of the Gaussian, whose σ is predefined.

(b) The probability density function is:

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

(c) The complexity for calculating a prediction is $O(nd)$, since we must perform the $x - \mu$ subtraction, which deals with d components, n times.

```
for each x_i where i = 1 to n:                    \\ O(n)
  calculate x_i - mu, which has d components       \\ O(d)
   square operation x_i - mu, which has d components    \\ O(d)
```

3. (a) The Parzen approach has a higher capacity/expressivity. Taking the sum over a set of kernels, where the kernels can be close together or far apart depending on the distribution of the data, allows for a highly nuanced density curve to emerge, while the parametric Gaussian approach is restricted to a density curve delimited in shape by the definition of a single Gaussian.

(b) The Parzen approach is also more likely to overfit and memorize noise, for the same reason. Its direct incorporation of the location of each datapoint as summands to the final density curve means that noise will contribute just as well to

the final density curve as the good data, while the parametric Gaussian approach generalizes this effect somewhat by using averages over the whole dataset and conforming to a less-nuanced shape.

(c) *Parameters* are learned by the algorithm from training data, whereas *hyperparameters* are assigned by the user, and validated on training and validation data. For example, in the parametric Gaussian density function questions below, we can optimize our μ parameter by finding the log error and taking derivatives, and this feedback process and learning of the parameter is the use case for the Gaussian modelling framework. By contrast, this Parzen framework involves the user observing the dataset and deciding what is the best σ value by "tuning" its value until a satisfactory result emerges. These types of algorithms must have users to set hyperparameters, because there is no efficient way of optimizing the function iteratively (e.g., in this case, the function is not convex).

4. Diagonal Gaussians

(a) To derive the equation of a diagonal Gaussian density with dimension d, we can start with the general Gaussian density:

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

where

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & 0 & \dots & 0 \\ 0 & \sigma_{ii}^2 & 0 & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & a_{dd}^2 \end{bmatrix}$$

and where

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_{11}^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_{ii}^2} & 0 & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{1}{\sigma_{dd}^2} \end{bmatrix}$$

Then, since the only filled positions in the Σ matrix are the diagonals, we can manually represent the matrix multiplication while skipping the calculations in the

matrix multiplication that do not interact with that diagonal:

$$p(x) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(\frac{-1}{2}\sum_{i=1}^{d}\frac{1}{\sigma_{ii}^2}(x_i - \mu_i)^2\right)$$

The above makes use of the fact that a transpose of a matrix symmetric about the diagonal is itself. Then, if we like, we can further represent the constant normalization term in product notation, using the fact that the determinant of a diagonal matrix is equal to the product of the diagonal terms:

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

The parameters of the diagonal Gaussian are:

- μ, the mean, of dimension d, and
- Σ, the covariance matrix, of dimension $d \times d$.

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

First, let us walk through the idea. We would like to prove the independence of the components of a diagonal Gaussian by proving that the product of each normal distribution's density function made up of 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:

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

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:

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

As above, we can also express the constant term in a product form, giving us the expression for $p(\boldsymbol{x})$:

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

Note that $\frac{1}{(2\pi\sigma_{ii}^2)^{1/2}} \exp\left(\frac{1}{2\sigma_{ii}^2}(x_i - \mu)^2\right)$ is a Gaussian with parameters $\mathcal{N}(\mu, \sigma_{ii})$. Thus, we have:

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

Therefore, the diagonal Gaussian density function is equal to the product of the individual components' density functions, and the individual components are independent. □

(c) Empirical risk can be written as:

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

Substituting from the previous part, we have:

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

(d) To find the optimal μ value, we take the derivative with respect to μ:

$$\begin{aligned}
\frac{\partial}{\partial\mu_i} &= \frac{1}{2|D|} \sum_{(x,y)\in D} \sum_{i=1}^{d} \frac{1}{\sigma_{ii}^2} 2(-x_i^{(j)} + \mu_i) \\
&= \frac{1}{|D|} \sum_{(x,y)\in D} \sum_{i=1}^{d} \frac{1}{\sigma_{ii}^2}(-x_i^{(j)} + \mu_i)
\end{aligned}$$

Setting $\frac{\partial}{\partial\mu_{ki}} = 0$, we have:

$$\begin{aligned}
\frac{1}{|D|} \sum_{(x,y)\in D} \sum_{i=1}^{d} \frac{1}{\sigma_{ii}^2}(-x_i^{(j)} + \mu_i) &= 0 \\
\sum_{(x,y)\in D} \sum_{i=1}^{d} \frac{1}{\sigma_{ii}^2} x_i &= \sum_{(x,y)\in D} \sum_{i=1}^{d} \frac{1}{\sigma_{ii}^2}\mu_i \\
\mu_i &= x_i^{(j)}
\end{aligned}$$

Therefore, the optimal μ value is $x_i^{(j)}$.

Now, to find the optimal σ^2 value, we take the derivative with respect to σ^2:

$$\begin{aligned}\frac{\partial}{\partial \sigma_{ii}^2} &= \frac{1}{2|D|} \sum_{(x,y)\in D} \sum_{i=1}^{d} \frac{1}{2\pi\sigma_{ii}^2} 2\pi - \frac{1}{\sigma_{ii}^4}(x_i^{(j)} + \mu_i) \\ &= \frac{1}{2|D|} \sum_{(x,y)\in D} \sum_{i=1}^{d} \frac{1}{\sigma_{ii}^2} - \frac{(x_i^{(j)} - \mu_i)}{\sigma_{ii}^4}\end{aligned}$$

Setting $\frac{\partial}{\partial \sigma_i^2}$ to zero, we have:

$$\begin{aligned}\frac{1}{2|D|} \sum_{(x,y)\in D} \sum_{i=1}^{d} \frac{1}{\sigma_{ii}^2} - \frac{(x_i^{(j)} - \mu_i)}{\sigma_{ii}^4} &= 0 \\ \frac{1}{2|D|} \sum_{(x,y)\in D} \sum_{i=1}^{d} \frac{1}{\sigma_{ii}^2} &= \frac{1}{2|D|} \sum_{(x,y)\in D} \sum_{i=1}^{d} \frac{(x_i^{(j)} - \mu_i)^2}{\sigma_{ii}^4} \\ \frac{1}{\sigma_{ii}^2} &= \frac{(x_i^{(j)} - \mu_i)^2}{\sigma_{ii}^4} \\ \sigma_{ii}^4 &= \sigma_{ii}^2 (x_i^{(j)} - \mu_i)^2 \\ \sigma_{ii}^2 &= (x_i^{(j)} - \mu_i)^2\end{aligned}$$

Therefore, the optimal σ^2 value is $(x_i - \mu_i)^2$.