DSC 255: Machine learning

Solutions to Homework 3

1. *Gaussian parameters.*

   (a) Based on the info given, we already know that $\mu = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$. Thus we just need to compute $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$.

$$\begin{aligned} \Sigma_{11} &= \text{var}(x) = \text{std}(x)^2 = 1 \\ \Sigma_{22} &= \text{var}(y) = \text{std}(y)^2 = \frac{1}{4} \\ \Sigma_{12} = \Sigma_{21} &= \text{cov}(x, y) = \text{corr}(x, y) \cdot \text{std}(x) \cdot \text{std}(y) = -\frac{1}{2} \cdot 1 \cdot \frac{1}{2} = -\frac{1}{4} \end{aligned}$$

   Thus $\Sigma = \begin{bmatrix} 1 & -0.25 \\ -0.25 & 0.25 \end{bmatrix}$.

   (b) We again can easily see $\mu = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. We can also write

$$\begin{aligned} \Sigma_{11} &= \text{var}(x) = \text{std}(x)^2 = 1 \\ \Sigma_{22} &= \text{var}(y) = \text{var}(x) = 1 \\ \Sigma_{12} = \Sigma_{21} &= \text{cov}(x, y) = \text{cov}(x, x) = \text{var}(x) = 1 \end{aligned}$$

   Thus $\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$.

2. This could easily happen if the training set contained a lot more $-$ points than $+$.

3. *Winery classification.*

   (a) 2
   (b) 2
   (c) 3
   (d) 1
   (e) 1

4. *Gaussian contours.* The contour lines should look something like this.

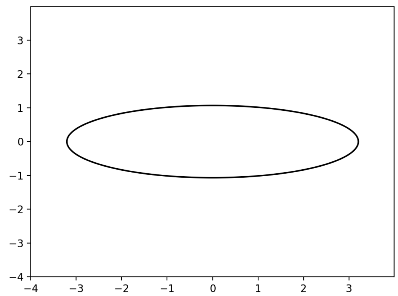


Figure 1: Contour line for (a)

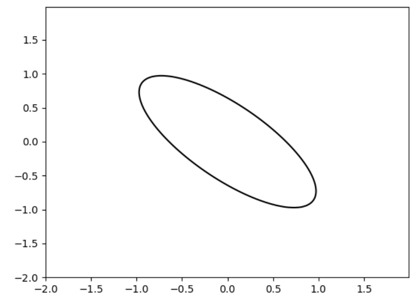


Figure 2: Contour line for (b)

5. $\begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$ and $\begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$

6. $x \cdot x = 25 \Leftrightarrow \|x\| = 5$. All points of length 5: a sphere, centered at the origin, of radius 5.

7. $f(x) = 2x_1 - x_2 + 6x_3 = w \cdot x$ for $w = \begin{bmatrix} 2 \\ -1 \\ 6 \end{bmatrix}$.

8. A is 10×30 and B is 30×20

9. (a) X is $n \times d$
   (b) XX^T is $n \times n$
   (c) $(XX^T)_{ij} = x^{(i)} \cdot x^{(j)}$

10. $x^T x = \|x\|^2 = 35$ and

$$x^T x = \begin{bmatrix} 1 & 3 & 5 \\ 3 & 9 & 15 \\ 5 & 15 & 25 \end{bmatrix}$$

11.

$$M = \begin{bmatrix} 3 & 1 & -2 \\ 1 & 0 & 0 \\ -2 & 0 & 6 \end{bmatrix}$$

12. (a) $|A| = 8! = 40320$
    (b) $A^{-1} = \text{diag}(1, 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8)$

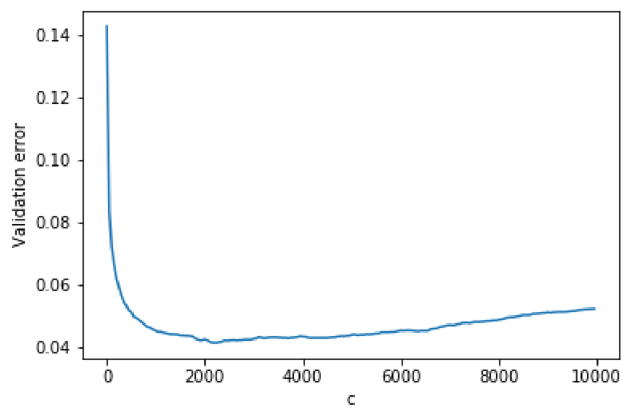
13. *Classifying MNIST using generative modeling.*

    (a) Pseudocode for training procedure:
    - Load in the original training data matrix X and label vector y.
    - Randomly split into validation set X_{valid}, y_{valid} of size 10,000 and training set X_{train}, y_{train}.
    - For each digit $i = 0, 1, 2, \ldots$:
      - Calculate fraction of data points in training set with label i: π_i
      - Calculate mean of data points in training set with label i: $\mu^{(i)}$

- Calculate covariance of data points in training set with label i: $\Sigma^{(i)}$
- For $c \in \{1, 51, 101, \ldots, 10001\}$:
  - Compute Gaussians $P_0 = \mathcal{N}(\mu^{(0)}, \Sigma^{(0)} + cI), \ldots, P_9 = \mathcal{N}(\mu^{(9)}, \Sigma^{(9)} + cI)$
  - Classify each validation point $x \in X_{\text{valid}}$ as the digit j which maximizes $\log \pi_j + \log P_j(x)$.
  - Compute the validation error (i.e. the fraction of validation points we misclassified).
- Select c^* to be the c which gave us the smallest validation error.

(b) We used a single value of c for all classes.

For a particular run, the above training procedure gives a validation error curve that looks like the following.

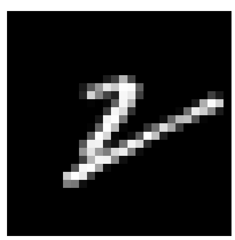


The c which achieves the minimum above is $c = 2151$. Note that your procedure may produce a different c due to the randomness in the choice of the validation/training split.

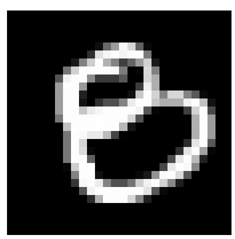
(c) The test error with this value is 0.0425.

(d) Now let's look at some randomly misclassified instances.

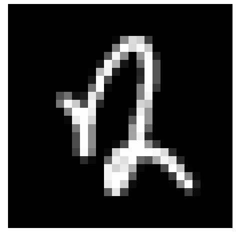
- The true label is 2, but it is predicted as 8.



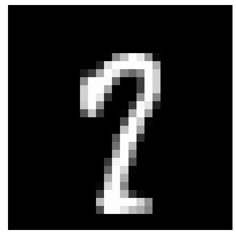
- The true label is 8, but it is predicted as 0.



- The true label is 2, but it is predicted as 4.



- The true label is 7, but it is predicted as 9.



- The true label is 6, but it is predicted as 0.