Course: DD2427 - Exercise Class 1

Questions with an asterix(*) are a bit more involved and are more to aid understanding as opposed to representing potentional exam questions.

Exercises 1: *Bayes I*

You have written a face detection algorithm. Let a denote the variable that there is a face in the image and b the output of your algorithm.

$$a = \begin{cases} 1 & \text{if there is a face in the image} \\ 0 & \text{there is not a face in the image} \end{cases} \qquad b = \begin{cases} 1 & \text{your algorithm reports there's a face in the image} \\ 0 & \text{your algorithm reports there's not a face in the image} \end{cases}$$

Your face detection algorithm has a false positive rate of .05 and a true positive rate of .85. Your algorithm is examining images that are taken from your front door.

You run your algorithm on an image taken at 10am (the time when the postman usually passes your house) and the result is positive. What is the probability the image contains a face ?

You run your algorithm on an image taken at 2am and the result is positive. What is the probability the image contains a face ?

Exercises 2: *Bayes Decision Theory*

A binary 2×2 image is generated by some random mechanism. By studying a large number of noise free realizations of the images generated it has been found that

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ has probability } \tfrac{1}{4},$$
$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ has probability } \tfrac{1}{4},$$
$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \text{ has probability } \tfrac{1}{2}$$

(a priori probabilities). One of these images has been distorted by noise in the sense that the value of a pixel has been changed with probability ϵ, that is

$$P(\text{observing } 0 \,|\, \text{the correct value is } 1) = P(\text{observing } 1 \,|\, \text{the correct value is } 0) = \epsilon$$

Assume that the noise in different pixels is independent. Now consider the

image

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

Using Bayes theorem calculate the MAP (maximum a posterior) estimation of the scene if

1. $\epsilon = 10\%$
2. $\epsilon = 50\%$

Exercises 3:

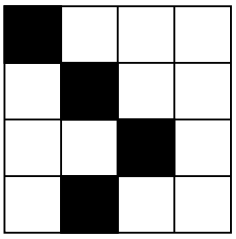
Consider a binary 4×4 image of a scene with a vertical line. In the *correct* image all pixels would be white except one vertical row with black pixels. Unfortunately, the camera used is far from perfect. Errors in different pixels are independent with

$$p(\text{white} \,|\, \text{line}) = p(\text{black} \,|\, \text{not line}) = \epsilon$$

and consequently

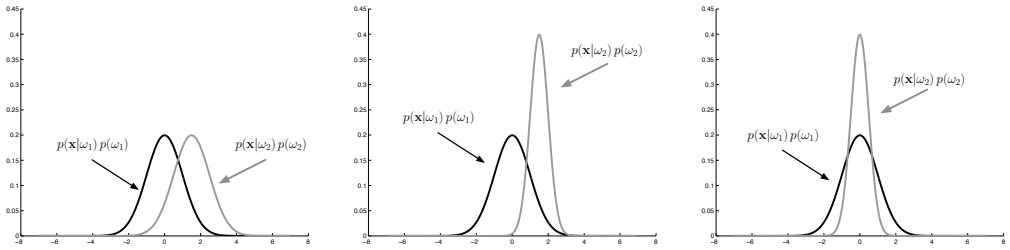
$$p(\text{black} \,|\, \text{line}) = p(\text{white} \,|\, \text{not line}) = 1 - \epsilon$$

Assume the a priori probability for the line to be located in column 1 or 4 is 0.3 (each) and the a priori probability that the line is in column 2 or 3 is 0.2 (each). Calculate the maximum a posteriori estimation of the following image when $\epsilon = 0.2$



Exercises 4: *Bayes' Decision Theory*

Assume you have a two class classification problem. Each class generates a one dimensional feature vector according to $p(x|\omega_i) = \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1, 2$. The prior probabilities for each class are $p(\omega_1) = p(\omega_2) = .5$. In the graphs below $p(x|\omega_i)\, p(\omega_i)$ for $i = 1, 2$ are shown for different values of the μ's and σ's. For each example $\mu_1 = 0, \sigma_1 = 1$ and then **1**) $\mu_2 = 1.5, \sigma_2 = 1$, **2**) $\mu_2 = 1.5, \sigma_2 = .5$ and **3**) $\mu_2 = 0, \sigma_2 = .5$



(**1**) $\mu_1 \neq \mu_2, \sigma_1 = \sigma_2$ (**2**) $\mu_1 \neq \mu_2, \sigma_1 > \sigma_2$ (**3**) $\mu_1 = \mu_2, \sigma_1 > \sigma_2$

i) For the two-class problem how is the *Bayes' Classifier* defined?

ii) In the figure draw the decision boundaries/boundary defined by a Bayes' classifier.

iii) For case (**2**) explicitly calculate the decision boundaries.

iv) For case (**2**) write down the $P(\text{error})$ for the *Bayes' Classifier* and show in a diagram where the errors are being made.

v) What is optimal about the *Bayes' Classifier*?

Exercises 5: *Bayes Risk I*

Consider the following 2 class classification problem. The likelihood functions for each class is a Gaussian:

$$P(x|\omega_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)$$

with $\mu_1 = 0, \sigma_1^2 = 5$ and $\mu_2 = 3, \sigma_2^2 = 1$. The priors for each class are $P(\omega_1) = P(\omega_2) = .5$. Define the (mis)classification costs as $C_{11} = C_{22} = 0$, $C_{12} = 1, C_{21} = \sqrt{5}$.

Determine a decision rule minimizing the probability of error.

Exercises 6: *k Nearest Neighbour classifier*

Remember the distance metric used in a nearest neighbour classifier affects the performance of the classifier. A commonly used distance metric family is the L_p norm where

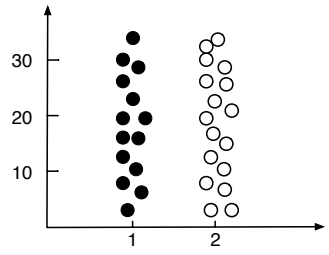
$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{d} |x_i|^p\right)^{\frac{1}{p}}$$

Consider the case of using a kNN classifier, but with the L_1 norm to measure distances rather than the L_2 (Euclidean) norm. Draw (in two dimensions) a simple case of a binary classification problem for which the L_1 classifier would return a different class for a test point than an L_2 classifier. In particular, draw ≥ 1 training points (one for each class) and a test point that would be classified differently according to the two distance metrics.

What properties of a data set do you imagine would influence whether the L_1 distance would work better or worse than the L_2 distance?

Exercises 7: *Nearest neighbour classification*

Carefully examine the data from two classes shown in the figure below.



Answer the following questions about this example.

i) Can you apply a kNN (say with $k = 5$) classifier on this data using a Euclidean distance metric and hope to obtain a sensible decision boundary? Explain your answer.

ii) How must the data be processed before a kNN will produce an accurate decision boundary ?

Exercises 8: *Nearest neighbour classification*

The bias of a classifier at a point $\mathbf{x}$ measures the amount by which the average of our estimate differs from the true class label:

$$\text{Bias} = \text{E}\left[L(y, \text{E}\left[\hat{f}(x)\right])^2\right]$$

while the variance of the classifier is the expected squared deviation of $\text{E}\left[\hat{f}(x)\right]$ around its mean

$$\text{Variance} = \text{E}\left[(\hat{f}(x) - \text{E}\left[\hat{f}(x)\right])^2\right]$$

Say we use the $0,1$ loss function and a kNN nearest classifier so that

$$f(\mathbf{x}) = \text{sgn}\left[\sum_{\mathbf{x}_i \text{ a neighbor of } \mathbf{x}} y_i\right]$$

what effect will the size of k have on the bias and variance of our classifier?

Exercises 9: *Discriminant Functions I*

Let $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \Sigma)$ for $i = 1, 2$ in a two-class d-dimensional problem with the same covariances but arbitrary means and prior probabilities.

a) Show that the decision boundary between the two classes is a hyper-plane.

b) Need this decision boundary be perpendicular to the line connecting the two means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

c) In terms of the prior probabilities for the two classes $P(\omega_1)$ and $P(\omega_2)$ state the condition that the Bayes decision does not pass between the two means.

Exercises 10: *Convergence of the linear perceptron learning rule*

Suppose we have n points $\mathbf{x}_i$ in $\mathcal{R}^p$ in general position, with class labels $y_i \in \{-1, 1\}$. Prove that the perceptron learning algorithm converges to a separating hyperplane in a finite number of steps by proving these sub-problems:

1. Denote a hyperplane by $f(\mathbf{x}) = \mathbf{w}_1^T \mathbf{x} + w_0 = 0$, or in more compact notation $\mathbf{w}^T \mathbf{x}^* = 0$, where $\mathbf{x}^* = (\mathbf{x}, 1)$ and $\mathbf{w} = (\mathbf{w}_1, w_0)$. Let $z_i = \mathbf{x}_i^*/\|\mathbf{x}_i^*\|$. Show that separability implies the existence of a $\mathbf{w}_{\text{sep}}$ such that $y_i \, \mathbf{w}_{\text{sep}}^T \, \mathbf{z}_i \geq 1 \ \ \forall i$

2. Given a current $\mathbf{w}_{\text{old}}$, the perceptron algorithm identifies a point $\mathbf{z}_i$ that is misclassified, and produces the update $\mathbf{w}_{\text{new}} \leftarrow \mathbf{w}_{\text{old}} + y_i \, \mathbf{z}_i$. Show that $\|\mathbf{w}_{\text{new}} - \mathbf{w}_{\text{sep}}\|^2 \leq \|\mathbf{w}_{\text{old}} - \mathbf{w}_{\text{sep}}\|^2 - 1$, and hence that the algorithm converges to a separating hyperplane in no more than $\|\mathbf{w}_{\text{start}} - \mathbf{w}_{\text{sep}}\|^2$ steps.

Exercises 11: *Cross Validation*

For k-fold cross-validation what are disdavantages and advantages for small and large k values? Why?

Exercises 12: *LDA*

Assume we have a two class problem. The feature vectors extracted from each class are two dimensional and the class conditional densities are:

$$p(\mathbf{x}|\omega_1) \sim N(\boldsymbol{\mu}_1, \Sigma) \quad \text{and} \quad p(\mathbf{x}|\omega_2) \sim N(\boldsymbol{\mu}_2, \Sigma)$$

where

$$\boldsymbol{\mu}_1 = (1,1)^T, \quad \boldsymbol{\mu}_2 = (3,2)^T \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 0.2^2 \end{pmatrix}$$

In LDA you project each feature vector generated by these class conditionals onto a line via $\mathbf{w}^T\mathbf{x}$ to obtain a scalar value.

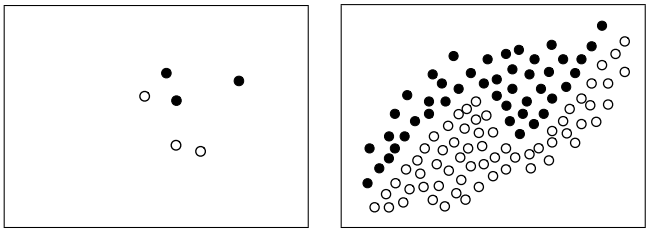
For $\sigma = 1$ sketch the class conditional densities and indicate why in this case it is better to project feature vectors from these two classes onto the y-axis as opposed to the x-axis for performing discrimination on the resulting scalar values.

For what values of σ will it be better to project onto the x-axis ?

For what values of σ will the y-axis tend towards the optimal projection line with respect to the Fisher criterion?

Exercises 13:

Draw the decision boundary formed by a 1-nearest neighbour classifier in the two different figures below. Not that the sparse set of points is a subset of the dense ones. What lesson should be learned from these examples?



(**a**) (**b**)

Exercises 14: *Naïve Bayes and Logistic regression**

Assume we have a two class problem. The feature vector $\mathbf{x} = (x_1, \ldots, x_d)$

extracted from each class is d dimensional and each $x_i \in \{0, 1\}$. Let

$$p(x_i = 1 \,|\, \omega = 1) = \theta_{i1} \quad p(x_i = 0 \,|\, \omega = 1) = 1 - \theta_{i1}$$
$$p(x_i = 1 \,|\, \omega = 0) = \theta_{i0} \quad p(x_i = 0 \,|\, \omega = 0) = 1 - \theta_{i0}$$

Show that

i) The above likelihoods can be written as $p(x_i \,|\, \omega = j) = \theta_{ij}^{x_i} \, (1 - \theta_{ij})^{1-x_i}$ for $j = 0, 1$.

ii) Assuming a independence between the features write down the expression for $p(\mathbf{x} \,|\, \omega = 0)$.

iii) If $P(\omega = 0) = p_0$ write down an expression for $p(\omega = 0 \,|\, \mathbf{x})$ (this corresponds to the Naive Bayes' model).

iv) Show how $p(\omega = 0 \,|\, \mathbf{x})$ can be written in the form

$$p(\omega = 0 \,|\, \mathbf{x}) = \frac{1}{1 + \exp\left(w_0 + \mathbf{w}^T \mathbf{x}\right)}$$

v) Then what are the expressions for $p(\omega = 0 \,|\, \mathbf{x})$ and

$$\log \frac{p(\omega = 1 \,|\, \mathbf{x})}{p(\omega = 0 \,|\, \mathbf{x})}$$

This is the same form as which discriminative model?

Exercises 15: *Fisher's Linear Discriminant and MSE**

The least squares approach to the determination of a linear discriminant was based on the goal of making the model predictions as close as possible to a set of target values. By constrast, the Fisher creterion is derived by requiring maximum class separation in the output space in conjuction with minimum within class spread. For the two-class problem the Fisher criterion can be seen as a special case of least squares.

Take the targets for class ω_1 and to be $\frac{n}{n_1}$ where n_1 is the number of patterns from class ω_1 and n is the total number of patterns. For class ω_2 take the targets to be $-\frac{n}{n_2}$ where n_2 is the number of patterns from class ω_2.

The sum-of-squares error function is written as

$$J = \frac{1}{2} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i + w_0 - t_i)^2$$

where each $t_i = \frac{n}{n_1}$ or $-\frac{n}{n_2}$ depending if $\mathbf{x}_i$ belongs to class ω_1 or ω_2. Show that J is minimized when

$$\mathbf{w} \propto S_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

where

$$\mathbf{m}_i = \frac{1}{n_i}\sum_{\mathbf{x}\in\omega_i}\mathbf{x} \quad \text{and} \quad S_{\mathbf{w}} = \sum_{\mathbf{x}\in\omega_1}(\mathbf{x}-\mathbf{m}_1)(\mathbf{x}-\mathbf{m}_1)^T + \sum_{\mathbf{x}\in\omega_2}(\mathbf{x}-\mathbf{m}_2)(\mathbf{x}-\mathbf{m}_2)^T$$

Exercises 16: *LDA II*

We have two class-conditional probabilities:

$$p(\mathbf{x}|\omega_1) \propto \exp\left(-.5\,\mathbf{x}^T\Sigma^{-1}\mathbf{x}\right)$$
$$p(\mathbf{x}|\omega_2) \propto \exp\left(-.5\,(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

where $\mathbf{x} = (x, y)$ is a two dimensional vector and

$$\boldsymbol{\mu} = (1,1)^T \qquad\qquad \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & \sigma^2 \end{pmatrix}$$

Define the function $f(\mathbf{x}) = ax + by$ which is the projection of the point $\mathbf{x}$ onto a line passing through the origin.

1. Compute the values $\mu_i = \mathrm{E}\left[f(\mathbf{x})|\omega_i\right]$ and $\sigma_i^2 = \mathrm{Var}\left[f(\mathbf{x})|\omega_i\right]$ for $i = 1, 2$
2. Find the values of a and b that maximize the statistical Fisher discriminant criterion:
$$J(a,b) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

Exercises 17: *Gaussians**

Have the following model for IQ and test scores S

$$IQ \sim \mathcal{N}(100, 15^2), \quad S \,|\, IQ \sim \mathcal{N}(IQ, 10^2)$$

You take the test and get a score of $s_1 = 130$.

i) Derive the posterior density $p(IQ \,|\, s_1 = 130)$

ii) You are a bit disappointed that Bayesians would consider you to have an IQ less then your test score. So you decide to take the test again. Let your score on the second test be S_2. What is the posterior desnity for $p(S_2 \,|\, s_1 = 130)$?