Course: DD2427 - Exercise Class 2

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

Exercises 1: *Linear Separability*

Given a set of data points $\{\mathbf{x}_i\}$, we can define the *convex hull* to be the set of all points $\mathbf{x}$ given by

$$\mathbf{x} = \sum_i \alpha_i \, \mathbf{x}_i$$

where $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$. Consider a set of points $\{\mathbf{y}_i\}$ together with their corresponding convex hull. By definition, the two sets of points will be linearly separable if there exists a vector $\hat{\mathbf{w}}$ and a scalar ω_0 such that $\hat{\mathbf{w}}^T \mathbf{x}_i + \omega_0 > 0$ for all $\mathbf{x}_i$, and $\hat{\mathbf{w}}^T \mathbf{x}_i + \omega_0 < 0$ for all $\mathbf{y}_i$. Show that if their convex hulls intersect, the two sets of points cannot be linearly separable, and conversely that if they are linearly separable, their convex hulls do not intersect.

Exercises 2: *Boosting*

Imagine you have started your ex-jobb project and you have hired some first year students to label some training data into two classes for you. Unfortunately, the night before working for you the student spent the night partying until early in the morning. Thus he has created a labelled dataset with lots of labelling errors, say upto 20% of the data is misclassified. What consequences will this have for your project if you are building a classifier using a boosting mechanism? (Or indeed using a nearest neighbour classifier?)

What learning mechanism could I use instead to fit separate models to the two different classes which would ignore the labels?

Exercises 3: *Boosting II*

a) What does the *AdaBoost* algorithm produce from a set of weak classifiers and labelled training data?

b) Describe the steps of the *AdaBoost* algorithm.

c) In your opinion what is the critical step in the *AdaBoost* algorithm?

d) What quality must the weak classifiers possess in order for *AdaBoost* to run successfully?

e) What are the strengths/weakness of using boosting to solve classification problems?

Exercises 4: Imagine you have a face detector such that

$$P(\hat{f} = 1 \,|\, f = 1) = 1 - \epsilon \quad \textbf{and} \quad P(\hat{f} = 0 \,|\, f = 0) = 1 - \alpha$$

where $f \in \{0, 1\}$ indicates the ground truth of whether a patch is present or not while $\hat{f} \in \{0, 1\}$ is the prediction of the face detector.

Imagine you have K *independent* face detectors each having the same true and false positive rate of the detector just described. If these detectors are applied to an image patch we get

$$\hat{\mathbf{f}} = (\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_K)$$

where $\hat{f}_i \in \{0, 1\}$ is the prediction of the ith detector. $k_{\hat{\mathbf{f}}} = \sum_i \hat{f}_i$ is equal to the number of detectors which predict a face while $K - k_{\hat{\mathbf{f}}}$ is the number which predict a non-face. Let γ be the prior probability that the patch contains a face. Given this information answer the following:

a) **(.2)** Write down the expression, remembering to exploit the independence, for

$$p(\hat{\mathbf{f}} \,|\, f = 1)$$

b) **(.2)** What is the posterior probability the patch contains a face given $\hat{\mathbf{f}}$?

c) **(.3)** Let $\gamma = \epsilon = \alpha = .01$ what is the constraint $k_{\hat{\mathbf{f}}}$ must fulfill such that

$$p(f = 1 \,|\, \hat{\mathbf{f}}) \geq .99$$

For $K = 4$, what is the minimal value of $k_{\hat{\mathbf{f}}}$ such that the above performance level is met? For $K = 10$? And as $K \to \infty$ what ratio of detections should correctly predict a face to ensure this level of performance.

d) **(.3)** Define a final classifier such that

$$F(\hat{\mathbf{f}}) = \begin{cases} 1 & \text{if } \sum_i \hat{f}_i \geq K_0 \\ 0 & \text{otherwise} \end{cases}$$

Continuing with the parameter settings just given, write down an expression for the $P(\text{error})$ of this classifier.

Exercises 5: *k means clustering*

Consider the k-means algorithm applied to a large amount of one-dimensional data that comes from either of one of two classes with equal prior probability. The class conditional distribution for each class is Gaussian with true means $\mu = \pm 1$ and both have standard deviation $\sigma = 1$. What happens when you apply the k-means algorithm with $k = 2$ to this data? What can you say about the means of the two clusters found and the mean of the class-conditional distributions.

Exercises 6: *Decision stumps*

The type of weak binary classifier for data $\mathbf{x} \in \mathbb{R}^d$ you have been using in the face lab has a name. It is called a *decision stump.* As you know the classification rule has parameters $q \in \{-1, 1\}, j \in \{1, 2, \ldots, d\}$ and θ and takes the form:

$$h(\mathbf{x}; j, q, \theta) = q \times \text{sgn}\,(x_j - \theta)$$

Decision stumps classify example $\mathbf{x}$ based only on the value of its jth coordinate. θ is a threshold value in $\mathbb{R}$ and q is the parity.

Decision stumps, by themselves, are not very powerful classifiers. For instance, a single vertical or horizontal decision stump can only shatter 2 points in $\mathbb{R}^2$. However, combining multiple decision stumps can give rise to more complex classifiers, as you do in the boosting algorithm. In this part, we calculate the VC dimensions of some combinations of decision stumps.

For points in R^2, calculate the VC-dimension of the following sets of classifiers:

a) Convex combinations (i.e. coefficients must be non-negative and sum to 1) of two vertical decision stumps.

b) Convex combinations of one vertical and one horizontal decision stump.

Exercises 7: *VC-dimension I*

Remember that in order to prove that a class of functions $\mathcal{H}$ has VC-dimension d you need to show that

- There exists a set of d points which can be **shattered** by $\mathcal{H}$.
- There exists **no** set of $d + 1$ points that can be shattered by $\mathcal{H}$

a) When does a class of functions $\mathcal{H}$ *shatter* a set of points $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$?

b) Show with appropriate diagrams that there exists 3 points in $\mathbb{R}^2$ that can be shattered by a line.

c) What is the VC-dimension of intervals in $\mathbb{R}$? In this case $\mathcal{H}$ is defined such that each $h \in \mathcal{H}$ is associated with an interval $[a, b]$ and $x \in \mathbb{R}$ has $h(x) = 1$ if and only if $x \in [a, b]$.

d) What is the VC-dimension of the union of k intervals on the real line? In other words each $h \in \mathcal{H}$ is associated with k closed intervals $[a_i, b_i]$, $i = 1, 2, \dots, k$ and $h(x) = 1$ if and only if $x \in \cup_{i=1}^{n} [a_i, b_i]$.

e) What is the VC-dimension of axis parallel rectangles in $\mathbb{R}^2$? In other words $h \in \mathcal{H}$ is associated with 2 closed intervals $[a_i, b_i]$ for $i = 1, 2$ and then for any $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$, $h(\mathbf{x}) = 1$ if and only if $x_i \in [a_i, b_i]$ for $i = 1, 2$.

f) Show that the VC-dimension of the class $\mathcal{H}$ of hyperplanes in $\mathbb{R}^2$ is 3?

g) Show that the VC-dimension of the class $\mathcal{H}$ of hyperplanes in $\mathbb{R}^d$ is $\geq d + 1$?

Exercises 8: *VC-dimension II*

Prove that a oriented hyper-plane cannot shatter $d + 2$ points in $\mathbb{R}^d$.

Exercises 9: *SVM*

i) Consider the degree-two polynomial kernel defined by $K(\mathbf{x}, \mathbf{z}) = (1 + \mathbf{x}^T\mathbf{z})^2$. Expand this out completely for the three-dimensional case (i.e., $\mathbf{x} = \langle x_1, x_2, x_3 \rangle$ and $\mathbf{z} = \langle z_1, z_2, z_3 \rangle$. Verify that this has the same form as the quadratic expansion, although with different coefficients on the terms.

ii) Continuing from the previous question, what is the form of Φ so that $K(\mathbf{x}, \mathbf{z}) = \Phi(\mathbf{x})^T\Phi(\mathbf{z})$? (You need only consider the three-dimensional data case.) How does this differ from the expansion

$$\Phi(\mathbf{x}) = \langle x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2, x_3 \rangle?$$

iii) Consider optimizing an SVM with *squared* loss on the ξ variables. That is, an optimization problem of the form:

$$\begin{aligned}&\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + \lambda \sum_n \xi_n^2 \text{ s.t.}\\ &\quad y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n \qquad (\forall n)\\ &\quad \xi_n \geq 0 \qquad (\forall n)\end{aligned}$$

Construct the dual formulation for this problem. In particular, construct the Lagrangian, optimize it with respect to $\mathbf{w}$ and b, plug these solutions back in and get an optimization problem just in terms of the dual (Lagrange) variables $\boldsymbol{\alpha}$. How does this compare to the dual formulation for the standard SVM?

iv) For D dimensional data, consider using the degree d polynomial kernel defined by $K(\mathbf{x}, \mathbf{z}) = (1 + \mathbf{x}^T\mathbf{z})^d$. What is the general form of the expansion? What are the coefficients on all the different forms in the expansion?

Exercises 10: *Mixture Models*

a) Sketch this one dimensional probability distribution

$$p(x) = \pi_1 \mathcal{N}(0, .5) + (1 - \pi_1)\mathcal{N}(5, 1)$$

when $\pi_1 = \frac{1}{2}$; $\pi_1 = .1$; **and** $\pi_1 = .9$.

b) If you have n points generated from $p(x)$ when $\pi_1 = \frac{1}{2}$ and you fit a Gaussian distribution to this data. Sketch what this distribution will look like. What's the problem here?

c) This issue highlights a problem with parametric methods. What is it?

d) What method is used to find the parameters of a Gaussian mixture model from training examples generated from the distribution?

Exercises 11: *Kernel magic**

Assume you are given m one dimensional training examples and their associated labels, that is $\{(x_i, y_i)\}_{i=1}^m$ where each $x_i \in \mathbb{R}^1$ and $y_i \in \{-1, +1\}$.

a) Draw a case where you have $m = 3$ training examples which are not linearly separable.

b) You know if you transform your one-dimensional data to a higher dimensional space then there is a higher likelihood that they will be linearly separable. Thus you define a feature transformation $\phi_n : \mathbb{R}^1 \to \mathbb{R}^n$ where

$$\phi_n(x) = \left(e^{-\frac{x^2}{2}}, x\, e^{-\frac{x^2}{2}}, \frac{x^2}{\sqrt{2}}\, e^{-\frac{x^2}{2}}, \ldots, \frac{x^n}{\sqrt{n!}}\, e^{-\frac{x^2}{2}} \right)$$

Explain why any set of 3 points (with no duplicates) can be linearly separated when transformed via ϕ_2. Similarly explain why any set of $n{+}1$ points (with no duplicates) can be linearly separated when transformed by ϕ_n.

c) Consider the case when $n \to \infty$ and ϕ_n becomes

$$\phi_\infty(x) = \left\{ e^{-\frac{x^2}{2}}, x\, e^{-\frac{x^2}{2}}, \frac{x^2}{\sqrt{2}}\, e^{-\frac{x^2}{2}}, \ldots, \frac{x^j}{\sqrt{j!}}\, e^{-\frac{x^2}{2}}, \ldots \right\}$$

Can you explicitly construct $\phi_\infty(x)$? (Not a trick question)

d) Is there a finite set of points, containing no duplicates, that cannot be linearly separated after applying ϕ_∞?

e) A linear classifier can be expressed using only the inner products of support vectors in the transformed feature space. The Kernel trick, exploited by the SVM, is to define a function $K(\cdot, \cdot)$ such that

$$K(x, y) = \phi_\infty(x) \cdot \phi_\infty(y)$$

where the inner product between two infinite vectors $\mathbf{a} = (a_1, a_2, \ldots)$ and $\mathbf{c} = (c_1, c_2, \ldots)$ is defined as

$$\mathbf{a} \cdot \mathbf{c} = \sum_{i=1}^{\infty} a_i\, b_i$$

Given the definition of ϕ_∞ compute the form of $K(x, y)$. Hint you may want to use the Taylor series expansion of e^x:

$$e^x = \lim_{n \to \infty} \sum_{j=0}^{n} \frac{x^j}{j!}$$

f) With such a high dimensional feature space should we be concerned about over-fitting?

Exercises 12: *EM**

We have two coins. The first is a fair coin while the second is not necessarily fair. In summary:

$$P(H|\text{coin } 1) = \frac{1}{2} \quad P(H|\text{coin } 2) = \alpha$$

This procedure is as follows:

> Coin 1 is tossed. If this results in a head then coin 1 is tossed again otherwise coin 2 is tossed.

a) What is the probability that the 2nd toss results in a head ?

b) The above process is repeated N independent times and n_2 times a head is obtained on the 2nd toss. What is the maximum likelihood estimate for α ?

c) Say we're told that the process was repeated N times and in total M heads were obtained (this includes the first and second toss). What two update equations can we repeatedly apply to obtain an estimate for α ?