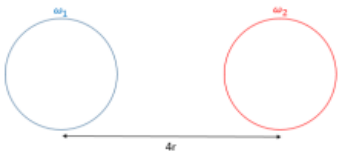
HW6

Q1:

When do we have error?



- If all N samples belong to ω_1 or all of them belong to ω_2 we have error in kNN ($k \geq 1$). Probability of having a sample from each class is $\frac{1}{2}$. Therefore, error in this case is $2(\frac{1}{2})^N = \frac{1}{2^{N-1}}$
- If N-1 samples belong to ω_1 and only one sample belong to ω_2, we will have no error for 1NN but we have error for kNN ($k \geq 3$): $\binom{N}{1}\left(\frac{1}{2}\right)^{N-1} \times (\frac{1}{2})$
- If N-1 samples belong to ω_2 and only one sample belong to ω_1, we will have no error for 1NN but we have error for kNN ($k \geq 3$): $\binom{N}{1}(\frac{1}{2}) \times \left(\frac{1}{2}\right)^{N-1}$
- If N-2 samples belong to ω_1 and only two samples belong to ω_2, we will have no error for 1NN and 3NN but we have error for kNN ($k \geq 5$): $\binom{N}{2}\left(\frac{1}{2}\right)^{N-2} \times (\frac{1}{2})^2$
- If N-2 samples belong to ω_2 and only two samples belong to ω_1, we will have no error for 1NN and 3NN but we have error for kNN ($k \geq 5$): $\binom{N}{2}\left(\frac{1}{2}\right)^{N-2} \times (\frac{1}{2})^2$
- ...
- In general,

$$p_e^{1NN} = \frac{1}{2^{N-1}}$$

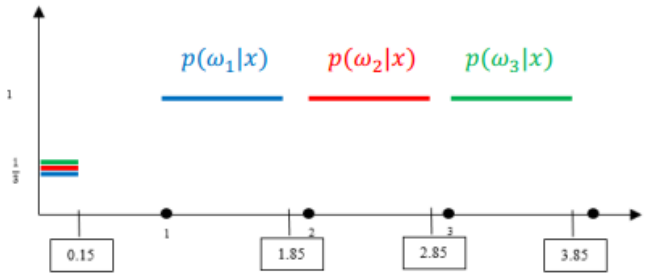
$$p_e^{3NN} = \frac{1}{2^{N-1}} + \binom{N}{1}\left(\frac{1}{2}\right)^{N-1}$$

$$p_e^{5NN} = \frac{1}{2^{N-1}} + \binom{N}{1}\left(\frac{1}{2}\right)^{N-1} + \binom{N}{2}\left(\frac{1}{2}\right)^{N-2} \times (\frac{1}{2})$$

...

So, $p_e^{kNN} > p_e^{1NN}$ for $k \geq 3$.

Q2: Part A)



$$p(x|\omega_i) = \begin{cases} 1 & 0 \le x \le 0.15 \\ 1 & i \le x \le i + 0.85 \\ 0 & O.W. \end{cases}$$

Using $p(\omega_i|x) = \frac{p(x|\omega_i)p(\omega_i)}{\sum_{i=1}^{3} p(x|\omega_i)p(\omega_i)}$:

$$p(\omega_1|x) = \begin{cases} \frac{1(\frac{1}{3})}{1(\frac{1}{3})+1(\frac{1}{3})+1(\frac{1}{3})} & 0 \le x \le 0.15 \\ \frac{1(\frac{1}{3})}{1(\frac{1}{3})} & 1 \le x \le 1.85 \\ 0 & O.W. \end{cases} = \begin{cases} \frac{1}{3} & 0 \le x \le 0.15 \\ 1 & 1 \le x \le 1.85 \\ 0 & O.W. \end{cases}$$

$$p(\omega_2|x) = \begin{cases} \frac{1}{3} & 0 \le x \le 0.15 \\ 1 & 2 \le x \le 2.85 \\ 0 & O.W. \end{cases} \qquad p(\omega_3|x) = \begin{cases} \frac{1}{3} & 0 \le x \le 0.15 \\ 1 & 3 \le x \le 3.85 \\ 0 & O.W. \end{cases}$$

Part B) We just speak about region [0 0.15] as Bayes classifier classifies the other parts with no error. Notice that Bayes classifier assign equal areas to each class in this region. Therefore,

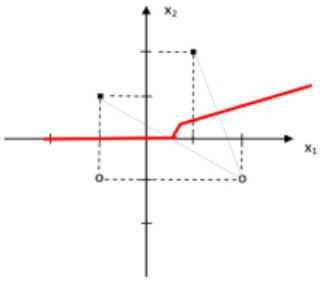
Bayes Error=$\left(\frac{1}{3}\int_0^{0.15} 1dx + \frac{1}{3}\int_0^{0.15} 1dx\right) = 0.1 = r$

We can get the same result by calculating parametrically for M classes.

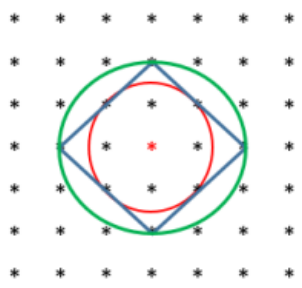
Part C) 1-NN error:

$$1 - \sum_{j=1}^{M} p^2(\omega_j|x) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{1}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = \frac{2}{3} \quad \rightarrow \quad \int_0^{0.15} \frac{2}{3} dx = 0.1 = r = \text{Bayes Error above}$$

Q3:



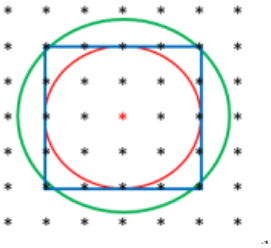
Q4: Part A) As we see in the picture below, The points on the blue line have the same Manhatan distance ($d_c = 2$).



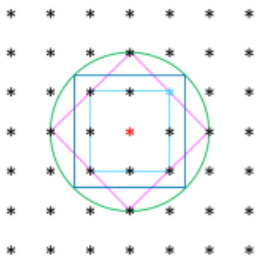
Suppose that it is the Manhatan distance of k[th] nearest sample to the test point x (red star). The Euclidean distance is always smaller than or equal to Manhatan distance. If we want to calculate the Euclidean distance of the k[th] nearest sample, at worst we get the points on the red circle ($d_e = \sqrt{2}$).

In set Y, we look for the points with the Manhatan distance smaller than $\sqrt{n}d_e$. In this case with 2 dimensions it becomes the green circle with r =2. As it is clear, The green circle contains all points of the blue set. All of this was about k[th] nearest neighbor. Therefore, the points nearer than k[th] nearest point to x are placed in the green circle.

Part B) It is similar to the algorithm above. We just need to put d_m instead of d_c. This time the radius of green circle is r =$2\sqrt{2}$ To have a better intuition compare and analyze this picture like above:



Part C&D) We should somehow merge two pictures in two parts above. For this aim consider the blue sample as the k[th] nearest neighbor of the red sample. With Manhatan distance the pink border is considered and according to part A we look for the points with Manhatan distance less than the green border, that is the pink border again.



With Maximum distance the light blue is considered and according to part B we must look for the points with maximum distance less than green circle, that is dark blue square.

Now we see the dark blue border and pink border both cover the same area although they do not have complete overlap. Minimum of these two distances can be proposed as a search region with smaller area.

Q5:

Part A)

$$\hat{p}(\mathbf{x}_0) = \frac{k_N}{N_n V_N} = \frac{k_N}{N_n h_N^n} = k_N \frac{1}{\sqrt[n]{N_n} h_N} \times \frac{1}{\sqrt[n]{N_n} h_N} \times \ldots \times \frac{1}{\sqrt[n]{N_n} h_N}$$

در رابطه فوق مقصود از N_n تعداد نمونه های مورد نیاز در فضای R^n می باشد. در دسته‌بند نزدیکترین همسایه داریم:

$$k_N = 1$$

و از آنجا:

$$\hat{p}(\mathbf{x}_0) \approx \prod_{i=1}^{n} \hat{p}(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt[n]{N_n} h_N}$$

با توجه به فرض الف، رابطه تخمین تابع چگالی احتمال در فضای یک‌بعدی خواهد بود:

$$\hat{p}(\boldsymbol{x}_0) = \frac{1}{N_1 h_N}$$

و از آنجا:

$$N_1 = \sqrt[n]{N_n} \Rightarrow N_n = N_1^n$$

$$N_1 = 100 \;\rightarrow\; N_n = 100^{20}$$

Part B)

حل: در فضای یک‌بعدی اگر N نمونه داشته باشیم و فرض کنیم بصورت یکنواخت در فاصله [0 1] توزیع شده باشند، آنگاه فاصله a بین هر دو نمونه خواهد بود:

$$a = \frac{1}{N}$$

در فضای R^n چنانچه فرض کنیم N نمونه بصورت کاملا یکنواخت و منظم در داخل فوق مکعب به ضلع واحد توزیع شده‌اند، آنگاه کوتاهترین فاصله بین دو نمونه مجاور خواهد بود:

$$a = \frac{1}{\sqrt[n]{N}}$$

و از آنجا با فرض N محدود داریم:

$$\lim_{n\to\infty} a = \lim_{n\to\infty} \frac{1}{\sqrt[n]{N}} = 1$$

مفهوم رابطه اخیر آن است که با فرض N محدود، فاصله بین نمونه‌ها به اندازه طول ضلع مکعب (که حاوی همه نمونه هاست) افزایش می‌یابد. یعنی نزدیکترین همسایه یک نمونه در فاصله بسیار زیاد از آن قرار می‌گیرد.

Part C)

حجم فوق مکعب به ضلع واحد: $V_1=1^n=1$

حجم فوق مکعب به ضلع $l_n(P)$: $V_{l_n(P)} = l_n^n(P)$

نسبت بین دو حجم:

$$\frac{V_{l_n(P)}}{V_1} = l_n^n(P) = P \Rightarrow l_n(P) = P^{1/n}$$

$l_5(0.01)=(0.01)^{1/5}=0.4$ $\quad$ $l_5(0.1)=(0.1)^{1/5}=0.63$

$l_{20}(0.01)=(0.01)^{1/20}=0.79$ $\quad$ $l_{20}(0.1)=(0.1)^{1/20}=0.89$

Part D)

$$d_M(\mathbf{x},\mathbf{x}') = \max_{i=1,\ldots,n}(x_i - x_i')$$

با توجه به نتیجه بند ب اگر فرض کنیم نمونه‌ها در داخل فوق مکعب بصورت یکنواخت توزیع شده باشند، آنگاه داریم:

$$E\{|x_i - x_i'|\} = N^{-1/n}$$

و از آنجا:

$$E\{d_M(\mathbf{x},\mathbf{x}')\} = N^{-1/n}$$

در این صورت مجموع فواصل $\mathbf{x}$ و $\mathbf{x}'$ از دو وجه فوق مکعب عمود بر راستای iام خواهد بود:

$$1 - N^{-1/n}$$

که با توجه به نتیجه بند ب:

$$\lim_{n\to\infty} 1 - \frac{1}{\sqrt[n]{N}} = 0$$

یعنی همه نقاط به یکی از وجوه فوق مکعب به ضلع واحد در فضای n بعدی نزدیکتر هستند.

Part E)

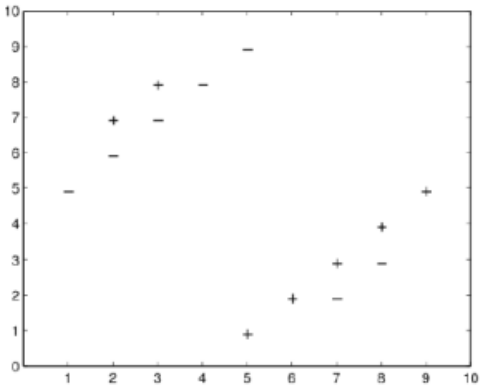
با استفاده از نتیجه فرض (د) می‌توان گفت که در فضا با ابعاد بالا نمونه‌ها عمدتا به سمت گوشه‌ها مهاجرت می‌کنند و لذا به دلیل پراکندگی نمونه‌ها در فضای ویژگی تخمین توزیع با استفاده از نمونه‌ها (که معمولا تعداد آنها در مقایسه با ابعاد بالای فضا کافی نیست) چندان دقیق نخواهد بود. وقتی با استفاده از مجموعه نمونه‌های آموزشی نتوان توزیع کلاسها را بخوبی تخمین زد، دقت کلاسیفایر حاصله نیز رضایتبخش نخواهد بود. بنابراین چنانچه N نمونه آموزشی در فضای n بعدی در اختیار داشته باشیم، هر چه بتوان ابعاد فضای ویژگی را با تحلیل صحیح داده‌ها کاهش داد و به ابعاد پایین‌تر n' ($n'<n$) رسید، دقت دسته‌بندی بالاتری مورد انتظار خواهد بود.

Q6:

Part A) class +

Part B) class –

Part C)



Part D) The best value for k is 5 that leads to 5NN. It has the least leave one-out cross validation error.

Q8:

Part A)

پاسخ: از آنجا که هیچگونه اطلاعی از توزیع کلاس‌ها در اختیار نیست، نمی توان از کلاسیفایر بیز استفاده کرد. زیرا در کلاسیفایر بیز باید فرم توزیع و پارامترهای آن از قبل در اختیار باشد.

با توجه به وجود تعداد قابل توجه نمونه از هر کلاس، استفاده از کلاسیفایرهای پارزن و kNN مناسب است ولی همانطور که می دانیم معمولا از روش پارزن برای تخمین توزیع و از KNN برای دسته‌بندی استفاده می‌شود.

بنابراین استفاده از روش kNN در این جا مناسب به نظر می‌رسد.

Part B)

پاسخ: در این شرایط تعداد نمونه ها نسبت به ابعاد فضای ویژگی کم است و با این تعداد نمونه، استفاده از پارزن یا kNN برای تخمین توزیع ها امکان پذیر نیست. در مقابل این تعداد نمونه در فضای ویژگی ۵۰ بعدی تا حد زیادی تضمین کننده تفکیک پذیری خطی آنها است. بنابراین کلاسیفایر خطی در اینجا دارای ترجیح است.

Part C)

پاسخ: روش kNN مستقیما احتمالات پسین کلاسها را تخمین می زند. زیرا داریم:

$$P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x}) \Rightarrow d(\mathbf{x}) = 1$$

با جایگذاری احتمالات پسین با احتمالات و احتمالات شرطی کلاس ها و حذف چگالی احتمال مرکب :

$$P(\omega_1)p(\mathbf{x}|\omega_1) > P(\omega_2)p(\mathbf{x}|\omega_2) \Rightarrow d(\mathbf{x}) = 1$$

با جایگذاری کمیت های فوق بوسیله تخمین های kNN آنها:

$$\frac{N_1}{N}\frac{k_1}{N_1V_N} > \frac{N_2}{N}\frac{k_2}{N_2V_N} \Rightarrow d(\mathbf{x}) = 1$$

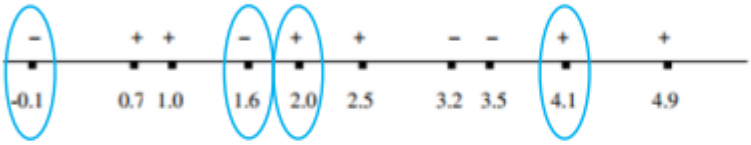
با ساده سازی:

$$k_1 > k_2 \Rightarrow d(\mathbf{x}) = 1$$

بنابراین k_1 و k_2 با تخمین احتمالات پسین کلاس ها متناسب هستند.

Q9:

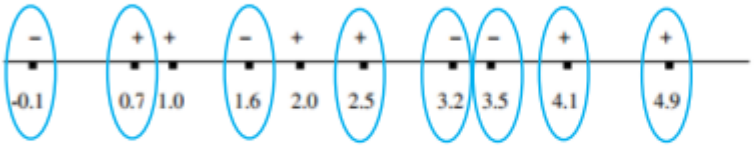
Part A)



Part B)

Leave-one-out cross-validation error is the number of misclassified items = $\frac{4}{10}$.

Part C)



Part D)

Leave-one-out cross-validation error is the number of misclassified items = $\frac{8}{10}$.

Q10:

Cosine distance: X

Euclidean distance: O