HW2 Intro to ML

---

1. (a)

$$\begin{aligned}
Z_y^{-1} &= \int \sum_{i:y_i=y} K(x, x_i)dx \\
&= \sum_{i:y_i=y} \int e^{-\frac{1}{\sigma^2}\|x-x_i\|^2} dx \\
&= \sum_{i:y_i=y} \int e^{-\frac{1}{\sigma^2}\|x\|^2} dx \\
&= \sqrt{\pi\sigma^2}\,|\{i|y_i = y\}|
\end{aligned}$$

for arbitrary $y \in \mathcal{Y}$. In turn, applying Bayes' Rule to the conditional (count) density functions, it holds

$$\begin{aligned}
\hat{y} = h(x) &= \arg \max_{y\in\{-1,1\}} \{f(y|x)\} \\
&= \arg \max_{y\in\{-1,1\}} \left\{ \frac{f(x|y)p(y)}{f(x)} \right\} \\
&= \arg \max_{y\in\{-1,1\}} \left\{ Z_y \left(\sum_{i:y_i=y} e^{-\frac{1}{\sigma^2}\|x-x_i\|^2} \right) \frac{1}{m}|\{i|y_i = y\}| \right\} \\
&= \arg \max_{y\in\{-1,1\}} \left\{ \frac{1}{m}(\sqrt{\pi\sigma^2}|\{i|y_i = y\}|)^{-1}|\{i|y_i = y\}| \left(\sum_{i:y_i=y} e^{-\frac{1}{\sigma^2}\|x-x_i\|^2} \right) \right\} \\
&= \arg \max_{y\in\{-1,1\}} \left\{ \frac{1}{m\sigma\sqrt{\pi}} \left(\sum_{i:y_i=y} e^{-\frac{1}{\sigma^2}\|x-x_i\|^2} \right) \right\} \\
&= \arg \max_{y\in\{-1,1\}} \left\{ \sum_{i:y_i=y} e^{-\frac{1}{\sigma^2}\|x-x_i\|^2} \right\} \\
&= \text{sign} \left(\sum_{i=1}^{m} y_i K(x, x_i) \right),
\end{aligned}$$

by using $K(x, x_i) > 0$ and $y_i \in \{-1, 1\}$.

(b) The Parzen-classifier returns the majority label among the m points, as the bandwidth parameter σ (that measures nearness of x to a data point x_i) is indifferent, considering each $\{x_1, \ldots, x_n\}$ equally "close" to the argument x. Therefore, the same weight is attributed to each data point (regardless of the actual distance $\rho(x, x_i)$). By dominated convergence (finite sums), it holds

$$\begin{aligned}
\lim_{\sigma\to\infty} f(x|y) &= \lim_{\sigma\to\infty} Z_y \left(\sum_{i:y_i=y} e^{-\frac{1}{\sigma^2}\|x-x_i\|^2} \right) \\
&= Z_y \sum_{i:y_i=y} \lim_{\sigma\to\infty} \left(e^{-\frac{1}{\sigma^2}\|x-x_i\|^2} \right) \\
&= Z_y \sum_{i:y_i=y} \mathbf{1} \\
&= Z_y \sum_{i=1}^{m} \mathbf{1}\{y_i = y\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\lim_{\sigma\to\infty} h(x) &= \lim_{\sigma\to\infty} \operatorname{sign}\left(\sum_{i=1}^{n} y_i K(x, x_i) \right) \\
&= \operatorname{sign}\left(\sum_{i=1}^{n} y_i \lim_{\sigma\to\infty} e^{-\frac{1}{\sigma^2}\|x-x_i\|^2} \right) \\
&= \operatorname{sign}\left(\sum_{i=1}^{m} y_i \right),
\end{aligned}$$

when allowing the limit and sign-function to commute. However, this assumption is non-trivial as the latter is non-continuous.

(c) This limit is more delicate, as $\lim_{\sigma\to 0} K(x, x_i)$ (Gaussian kernel) is singular. However, one can consider this limit belonging to the degenerate (point) distribution $\delta_{x_i}(x)$. In the limit, the Parzen classifier is "hyper-local", i.e. it memorized the training data and predicts the (arbitrary) default label $y_0 \in \mathcal{Y}$ for $x \notin \{x_1, \ldots, x_m\}$. Therefore, it is unable to generalize.
It holds

$$\begin{aligned}
\lim_{\sigma\to 0} h(x) &= \operatorname{sign}\left(\sum_{i=1}^{m} y_i \lim_{\sigma\to 0} e^{-\frac{1}{\sigma^2}\|x-x_i\|^2} \right) \\
&= \operatorname{sign}\left(\sum_{i=1}^{m} y_i \mathbf{1}\{x = x_i\} \right) \\
&= \begin{cases} 1, & x_i = x, \\ 0, & \text{else}. \end{cases}
\end{aligned}$$

If, on the other hand, ties are present coinciding with x, i.e. $x_i = \ldots = x_k = x$ for several pairwise different $\{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$, a "hyper-local" majority rule is performed among those points that are tied.

(d) It holds $\rho(x, x^{(1)}) \gg \rho(x, x^{(2)}) > \ldots > \rho(x, x^{(m)})$ with $x^{(i)}$ denoting the i-nearest data point to x. Further, let $y^{(i)}$ denote the corresponding label to $x^{(i)}$. Then, it holds

$$\begin{aligned}
h(x) = \operatorname{sign}\left(\sum_{i=1}^{m} y_i K(x, x_i) \right) &= \operatorname{sign}\left(\sum_{i=1}^{m} y_i e^{-\frac{1}{\sigma^2}\|x-x_i\|^2} \right) \\
= \operatorname{sign}\left(y^{(1)} e^{-\frac{\rho(x,x^{(1)})^2}{\sigma^2}} + y^{(2)} e^{-\frac{\rho(x,x^{(2)})^2}{\sigma^2}} + \ldots + y^{(m)} e^{-\frac{\rho(x,x^{(m)})^2}{\sigma^2}} \right)
\end{aligned}$$

$$= \operatorname{sign}\left(y^{(1)} \left(e^{-\sigma^{-2}}\right)^{\rho(x,x^{(1)})^2} + y^{(2)} \left(e^{-\sigma^{-2}}\right)^{\rho(x,x^{(2)})^2} + \ldots + y^{(m)} \left(e^{-\sigma^{-2}}\right)^{\rho(x,x^{(m)})^2} \right)$$

$$\approx \operatorname{sign}(y^{(1)}) = y^{(1)} = y_{\underset{i}{\operatorname{argmin}} \rho(x,x^{(i)})} = h_{NN}(x).$$

Therefore, the Parzen classifier h behaves approximately like h_{NN} if σ is sufficiently large and $x^{(1)}$ comparatively close to x (as compared to $x^{(2)}$ etc.). In case of ties, this Parzen classifier implicitly performs majority rule while the Nearest Neighbor classifier breaks a tie by arbitrarily picking one label. (In case of k tied points, the Parzen classifier would coincide with a kNN classifier. However, if X is continuous, the probability of such an event is 0.)

2. (a) It holds

$$h(x) = h_{\text{Bayes}}(x) = \arg\min_h L_{D(h)}(h) = \arg\min_h \mathbb{P}_D(h(x) \neq y).$$

Therefore,

$$h_{\text{Bayes}}(x) = \operatorname{sign}(x)$$

as it minimizes the conditional probability of an erroneous prediction for $x \geq 0$ and $x < 0$ to 0.2, respectively.

$$\begin{aligned}
\mathbb{P}_D(h_{\text{Bayes}}(x) \neq y) &= \mathbb{P}_D(Y = -1, \operatorname{sign}(X) = 1 | X \geq 0)\mathbb{P}_D(X \geq 0) + \mathbb{P}_D(Y = 1, \operatorname{sign}(X) = -1 | X < 0)\mathbb{P}_D(X < 0) \\
&= 0.2 \cdot 0.5 + 0.2 \cdot 0.5 \\
&= 0.2.
\end{aligned}$$

In short, $L_D(h_{\text{Bayes}}) = 0.2$.

(b) In the following, the notation is used. For some $x \in \mathbb{R}$ (likely contained chosen in $[-1, 1]$) by the linearity and tower property of the conditional expectation operator, it holds

$$\begin{aligned}
L_D(h_m) &= \mathbb{P}_D(h_m \neq y) \\
&= \mathbb{P}_D(h_m(x) = 1, Y = -1) + \mathbb{P}_D(h_m(x) = -1, Y = 1) \\
&= \mathbb{P}_D(X_N N(Y) = 1, Y = -1) + \mathbb{P}_D(X_N N(Y) = -1, Y = 1) \\
&= \mathbb{E}_X \left[\mathbb{P}_D(X_N N(Y) = 1, Y = -1 | X \geq 0) + \mathbb{P}_D(X_N N(Y) = -1, Y = 1 | X < 0)\right] \\
&= \mathbb{E}_X \left[\mathbb{P}_D(Y = -1 | X \geq 0)\mathbb{P}_D(X_N N(Y) = 1 | X \geq 0) + \mathbb{P}_D(Y = 1 | X < 0)\mathbb{P}_D(X_N N(Y) = -1 | X < 0)\right] \\
&= \mathbb{E}_X \left[(0.3 + \operatorname{sign}(X))(0.3 - \operatorname{sign}(X)) + (0.3 - \operatorname{sign}(X))(0.3 + \operatorname{sign}(X))\right] \\
&= \frac{1}{2}(0.8 \cdot 0.2 + 0.8 \cdot 0.2) + \frac{1}{2}(0.8 \cdot 0.2 + 0.8 \cdot 0.2) \\
&= 0.32.
\end{aligned}$$

3. (a) We can shatter four points as follows (excluding some cases here due to symmetry)

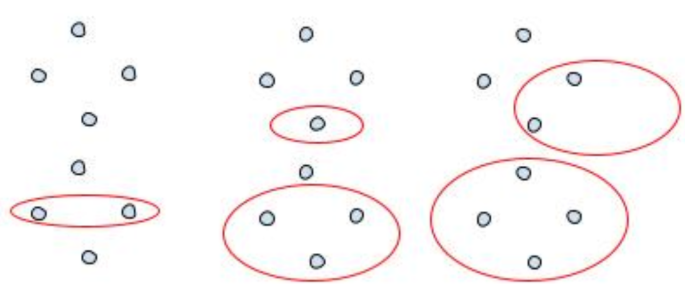

Figure 1:

(b) If we expand out the equation for the hypothesis class we get:

$$a_2^2(x_1 - c_1)^2 + a_1^2(x_2 - c_2)^2 \leq r^2 a_1^2 a_2^2$$

$$a_2^2 x_1^2 - 2a_2^2 c_1 x_1 + a_2^2 c_1^2 + a_1^2 x_2^2 - 2a_1^2 c_2 x_2 + a_1^2 c_2^2 \leq r^2 a_1^2 a_2^2$$

$$a_2^2 x_1^2 - a_2^2 2c_1 x_1 + a_2^2 c_1^2 + a_1^2 x_2^2 - a_1^2 2c_2 x_2 + a_1^2 c_2^2 - r^2 a_1^2 a_2^2 \leq 0$$

So we see we have a linear combination of the terms x_1^2, x_2^2, x_1, x_2 and a constant. So, we define ϕ as:

$$\phi : (x_1, x_2) \mapsto (x_1^2, x_2^2, x_1, x_2, 1)$$

(c) The VC dimension is 5. We can clearly shatter a rotated pentagon, but we cannot shatter 6 or more points. For more than 6 points, there are two cases:

i. If there is some point that is not on the convex hull of the points, we cannot shatter only the points in the convex hull.

ii. If all the points are in the convex hull, we can select every other point from the convex hull, which cannot be shattered separate from the other points.

In (a), we showed that we can shatter 4 points. This is consistent with the VC dimension here since the VC dimension being 5 implies we can shatter 4 points.

4. (a) For $n < \log(d)$ points, there are at most d possible labelings for them. So, for the ith such labeling let's make the ith coordinate of each point $+1$ or -1 according to that labeling. Now, we can shatter these points by taking $\mathbf{w}$ as the vector with everything zero except coordinate i which is 1 to get the ith labeling of the points. This is $\Omega(\log(d))$, as required.

(b) We already know how to shatter $\log(d)$ points. If $k > \log(d)$, then we know the VC dimension of hyperplanes without a bias term in k dimensions is exactly k so we can use that to shatter k points. Combining this tells us we can shatter $\max(\log(d), k)$ points.