Solutions to selected exercises from Chapter 22 of *Wasserman — All of Statistics*

(3) See the Jupyter Notebook 3.ipynb.

(4) The VC-dimension of $\mathcal{A}$ is 3. One may check that any set of three non-collinear points is shattered by $\mathcal{A}$.

So we consider four points in $\mathbb{R}^2$ and we claim they cannot be shattered. If three of the points, say $x, y, z$ are collinear, then one of them, say $y$, lies on the line segment $[x, z]$. On the other hand, if $A \in \mathcal{A}$ is a disk containing $x, z$ then it also contains $[x, z]$ by convexity and hence it contains $y$. Otherwise we have four points $x, y, z, w$ in general position. The set of line segments between points in $\{x, y, z, w\}$ is a quadrilateral with line segments joining its two pairs of opposite vertices. Thus two such line segments cross. Say $[x, y]$ and $[z, w]$ cross. Suppose there is a disk $A$ containing $\{x, y\}$ but not $\{z, w\}$ and a disk $B$ containing $\{z, w\}$ but not $\{x, y\}$. The intersection of $A$ and $B$ is a lens $L$ bounded by an arc of the boundary $\partial A$ and an arc of the boundary $\partial B$. Denote the two points of intersection of these arcs by $p$ and $q$. Denote the bi-infinite line through $p$ and $q$ by $\mathcal{L}$. Then one side of $\mathcal{L}$ contains no point of $A \setminus B$ and the other side contains no point of $B \setminus A$. Hence $[x, y]$ lies on one side of $\mathcal{L}$ and $[z, w]$ lies on the other side so that $[x, y]$ and $[z, w]$ cannot cross. This is a contradiction.

(5) See the Jupyter Notebook 5.ipynb.

(6) See the Jupyter Notebook 6.ipynb.

(7) It is clear that no linear classifier can perfectly classify the data assuming there are some $i$ falling into the three different cases $X_i < -1$, $-1 \leq X_i \leq 1$, and $X_i > 1$. On the other hand, the data $Z_i$ can be separated by the plane $y = 1$ in $\mathbb{R}^2$.

(8) See the Jupyter Notebook 8.ipynb.

(9) Apply the $k$ nearest neighbors classifier to the "iris data." Choose $k$ by cross-validation.

(10) This is actually the formula for the median distance for $n$ points in the *unit ball*. See e.g. Hastie-Tibshirani-Friedman equation (2.24). Moreover, the correct expression for the median is actually $(1 - (1/2)^{1/n})^{1/d}$ (i.e. there is no need for the factor $v_d(1)^{-1/d}$).

To prove this equation for the unit ball, note that if $X_1, \ldots, X_n$ are uniformly distributed, then for any $r \in [0, 1]$

$$\mathbb{P}(R > r) = \mathbb{P}(|X_i| > r \text{ for all } i).$$

The volume of the $r$-ball is $v_d(r) = r^d v_d(1)$, and renormalizing to give the unit ball volume 1 by dividing by $v_d(1)$, the volume is just $r^d$. Therefore

$$\mathbb{P}(R > r) = 1 - (1 - r^d)^n \text{ and } \mathbb{P}(R \leq r) = (1 - r^d)^n.$$

The median is given by solving

$$(1 - r^d)^n = \frac{1}{2}$$

which yields $r = \left(1 - \frac{1}{2^{1/n}}\right)^{1/d}$ as claimed.

I'm not sure what the correct expression is for the median closest distance for the cube $[-1/2, 1/2]^d$, but it seems pretty tedious to calculate.

(11) See the Jupyter Notebook 11.ipynb.

(12) See the Jupyter Notebook 12.ipynb.