DSC 255: Machine learning

Solutions to HW2

1. (a) $\|x - x'\|_2 = \sqrt{8}$
   (b) $\|x - x'\|_1 = 4$
   (c) $\|x - x'\|_\infty = 2$

2. (a) $\|x\|_1 = 6$
   (b) $\|x\|_2 = \sqrt{14}$
   (c) $\|x\|_\infty = 3$

3. The distance function is **not** a metric. Let's consider the four metric properties:

   - $d(x, y) \geq 0$: satisfied
   - $d(x, y) = 0$ if and only if $x = y$: satisfied
   - $d(x, y) = d(y, x)$: satisfied
   - Triangle inequality: not satisfied. E.g. $d(A, D) > d(A, C) + d(C, D)$.

4. The KL divergence between the two distributions is

$$\begin{aligned} K(p, q) &= \frac{1}{2} \log \frac{1/2}{1/4} + \frac{1}{4} \log \frac{1/4}{1/4} + \frac{1}{8} \log \frac{1/8}{1/6} + \frac{1}{16} \log \frac{1/16}{1/6} + \frac{1}{16} \log \frac{1/16}{1/6} \\ &= \frac{1}{2} \log 2 + \frac{1}{4} \log 1 + \frac{1}{8} \log \frac{3}{4} + \frac{1}{16} \log \frac{3}{8} + \frac{1}{16} \log \frac{3}{8} \ \approx \ 0.27 \end{aligned}$$

5. *Classification or regression?*

   (a) Classification
   (b) Regression
   (c) Regression
   (d) Classification

6. *Variance examples.*

   (a) Variance 1.
   Explanation: X takes value -1 with probability $1/2$ and value 1 with probability $1/2$. Therefore $\mathbb{E}[X] = 0$. Now, $X^2 = 1$ always, so $\mathbb{E}[X^2] = 1$. Hence $\text{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = 1$.
   (b) Variance 0.
   Explanation: X does not vary at all, so $X = \mathbb{E}[X]$ always. Thus $\text{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = 0$.
   (c) Variance 3/8.
   Explanation: X is a coin of bias $p = 1/4$. Its variance is $p(1-p) = 1/4 \times 3/4 = 3/16$.

7. *Covariance and correlation.*

(a) $\text{cov}(X, Y) = -2/3$.
Explanation: First, let's determine individual statistics for X and Y.
X has the following distribution:

x	$\Pr(X = x)$
-1	$1/3$
0	$1/3$
1	$1/3$

Thus $\mathbb{E}[X] = 0$ and $\text{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = (1/3)1^2 + (1/3)0^2 + (1/3)(-1)^2 = 2/3$.
Next, let's look at the distribution of Y.

y	$\Pr(Y = y)$
-1	$1/3$
0	$1/3$
1	$1/3$

This is the same as that of X, so $\mathbb{E}[Y] = 0$ and $\text{var}(Y) = 2/3$.
Finally, let's look at the product XY.

xy	Pr
-1	$2/3$
0	$1/3$
1	0

Thus $\mathbb{E}[XY] = -2/3$ and $\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = -2/3$.

(b) $\text{corr}(X, Y) = -1$
Explanation: We have already seen that $\text{var}(X) = \text{var}(Y) = 2/3$. Thus

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\text{std}(X)\text{std}(Y)} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{-2/3}{\sqrt{(2/3)(2/3)}} = -1.$$

8. *Independence and uncorrelatedness.*

(a) Not independent.
Explanation: Notice that $\Pr(X = 0) = 1/3$ and $\Pr(Y = 0) = 1/3$, but $\Pr(X = 0, Y = 0) = 0$.

(b) Uncorrelated.
Explanation: X takes on values -1, 0, or 1, each with probability 1/3. Therefore $\mathbb{E}[X] = 0$. The same is true of Y and of XY. Thus $\mathbb{E}[Y] = 0$ and $\mathbb{E}[XY] = 0$, whereupon $\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$.

9. *Classifying back injuries.* The code can be found in the accompanying notebook `back-injuries.ipynb`.

(a) Error rate with l_1 distance = 21.67%
Error rate with l_2 distance = 23.33%

(b) Confusion matrix for l_1 distance:

	NO	DH	SL
NO	14	0	2
DH	9	9	0
SL	1	1	24

Confusion matrix for l_2 distance:

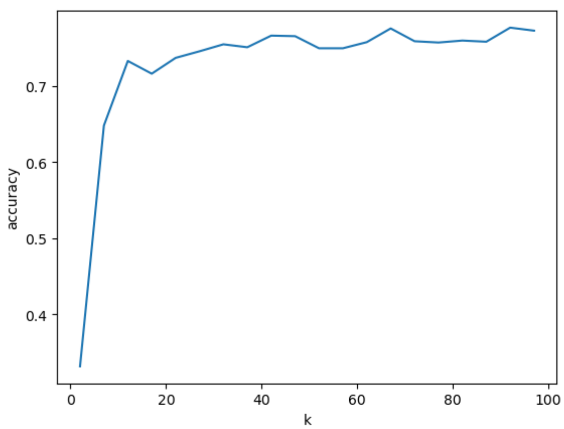
	NO	DH	SL
NO	12	1	3
DH	9	9	0
SL	1	0	25

10. *Cross-validation for nearest neighbor classification.* Code is given in the accompanying notebook `wine-nn-crossval.ipynb`.

    (a) Using leave-one-out cross-validation, we get the following estimates: the accuracy is ≈ 0.77 and the confusion matrix is

    $$\begin{bmatrix} 52 & 3 & 4 \\ 5 & 54 & 12 \\ 3 & 14 & 31 \end{bmatrix}$$

    (b) Here is the plot we get:



    (c) We normalize the data by linearly remapping each feature to have minimum value 0 and maximum value 1. After this normalization, the leave-one-out estimate of accuracy is ≈ 0.95 and the estimated confusion matrix is

    $$\begin{bmatrix} 59 & 0 & 0 \\ 5 & 62 & 4 \\ 0 & 0 & 48 \end{bmatrix}$$

    In this instance, normalization helps a lot!