CS 188 Fall 2020 Introduction to Artificial Intelligence

Exam Prep 10 Solutions

Q1. Naive Bayes: Pacman or Ghost?

You are standing by an exit as either Pacmen or ghosts come out of it. Every time someone comes out, you get two observations: a visual one and an auditory one, denoted by the random variables X_v and X_a, respectively. The visual observation informs you that the individual is either a Pacman ($X_v = 1$) or a ghost ($X_v = 0$). The auditory observation X_a is defined analogously. Your observations are a noisy measurement of the individual's true type, which is denoted by Y. After the indiviual comes out, you find out what they really are: either a Pacman ($Y = 1$) or a ghost ($Y = 0$). You have logged your observations and the true types of the first 20 individuals:

individual i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
first observation $X_v^{(i)}$	0	0	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	0	0	0
second observation $X_a^{(i)}$	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
individual's type $Y^{(i)}$	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0

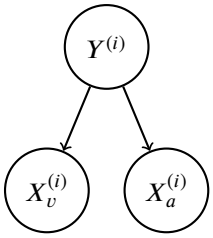
The superscript (i) denotes that the datum is the ith one. Now, the individual with $i = 20$ comes out, and you want to predict the individual's type $Y^{(20)}$ given that you observed $X_v^{(20)} = 1$ and $X_a^{(20)} = 1$.

(a) Assume that the types are independent, and that the observations are independent conditioned on the type. You can model this using naïve Bayes, with $X_v^{(i)}$ and $X_a^{(i)}$ as the features and $Y^{(i)}$ as the labels. Assume the probability distributions take on the following form:

$$P(X_v^{(i)} = x_v | Y^{(i)} = y) = \begin{cases} p_v & \text{if } x_v = y \\ 1 - p_v & \text{if } x_v \neq y \end{cases}$$

$$P(X_a^{(i)} = x_a | Y^{(i)} = y) = \begin{cases} p_a & \text{if } x_a = y \\ 1 - p_a & \text{if } x_a \neq y \end{cases}$$

$$P(Y^{(i)} = 1) = q$$



for $p_v, p_a, q \in [0, 1]$ and $i \in \mathbb{N}$.

(i) What's the maximum likelihood estimate of p_v, p_a and q?

$p_v =$ $\frac{4}{5}$ $p_a =$ $\frac{3}{5}$ $q =$ $\frac{1}{2}$

To estimate q, we count 10 $Y = 1$ and 10 $Y = 0$ in the data. For p_v, we have $p_v = 8/10$ cases where $X_v = 1$ given $Y = 1$ and $1 - p_v = 2/10$ cases where $X_v = 1$ given $Y = 0$. So $p_v = 4/5$. For p_a, we have $p_a = 2/10$ cases where $X_a = 1$ given $Y = 1$ and $1 - p_v = 0/10$ cases where $X_v = 1$ given $Y = 0$. The average of $2/10$ and 1 is $3/5$.

(ii) What is the probability that the next individual is Pacman given your observations? Express your answer in terms of the parameters p_v, p_a and q (you might not need all of them).

$$P(Y^{(20)} = 1 | X_v^{(20)} = 1, X_a^{(20)} = 1) = \frac{p_v p_a q}{p_v p_a q + (1-p_v)(1-p_a)(1-q)}$$

The joint distribution $P(Y = 1, X_v = 1, X_a = 1) = p_v p_a q$. For the denominator, we need to sum out over Y, that is, we need $P(Y = 1, X_v = 1, X_a = 1) + P(Y = 0, X_v = 1, X_a = 1)$.

Now, assume that you are given additional information: you are told that the individuals are actually coming out of a bus that just arrived, and each bus carries *exactly* 9 individuals. Unlike before, the types of every 9 consecutive individuals are *conditionally* independent given the bus type, which is denoted by Z. Only after all of the 9 individuals have walked out, you find out the bus type: one that carries mostly Pacmans ($Z = 1$) or one that carries mostly ghosts ($Z = 0$). Thus, you only know the bus type in which the first 18 individuals came in:

individual i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
first observation $X_v^{(i)}$	0	0	1	0	1	0	0	1	1	1	0	1	1	0	1	1	1	0	0	0
second observation $X_a^{(i)}$	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
individual's type $Y^{(i)}$	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0

bus j	0	1	
bus type $Z^{(j)}$	0	1	

(b) You can model this using a variant of naïve bayes, where now 9 consecutive labels $Y^{(i)}, \ldots, Y^{(i+8)}$ are *conditionally* independent given the bus type $Z^{(j)}$, for bus j and individual $i = 9j$. Assume the probability distributions take on the following form:

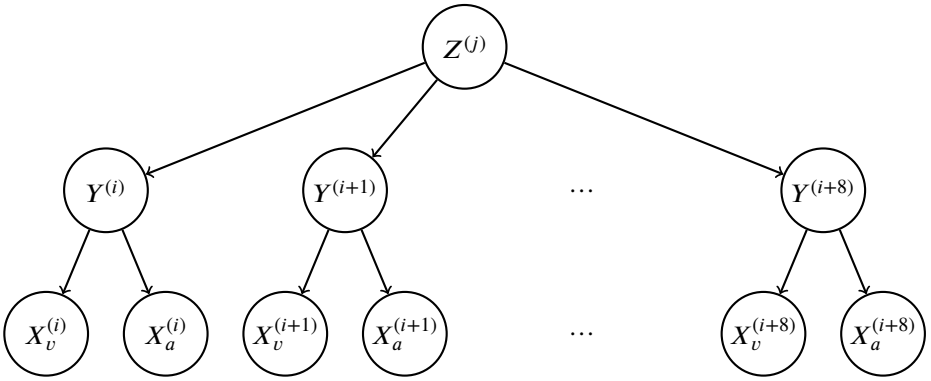
$$P(X_v^{(i)} = x_v | Y^{(i)} = y) = \begin{cases} p_v & \text{if } x_v = y \\ 1 - p_v & \text{if } x_v \neq y \end{cases}$$

$$P(X_a^{(i)} = x_a | Y^{(i)} = y) = \begin{cases} p_a & \text{if } x_a = y \\ 1 - p_a & \text{if } x_a \neq y \end{cases}$$

$$P(Y^{(i)} = 1 | Z^{(j)} = z) = \begin{cases} q_0 & \text{if } z = 0 \\ q_1 & \text{if } z = 1 \end{cases}$$

$$P(Z^{(j)} = 1) = r$$

for $p, q_0, q_1, r \in [0, 1]$ and $i, j \in \mathbb{N}$.



(i) What's the maximum likelihood estimate of q_0, q_1 and r?

$q_0 =$ $\frac{2}{9}$ $q_1 =$ $\frac{8}{9}$ $r =$ $\frac{1}{2}$

For r, we've seen one ghost bus and one pacman bus, so $r = 1/2$. For q_0, we're finding $P(Y = 1 | Z = 0)$, which is $2/9$. For q_1, we're finding $P(Y = 1 | Z = 1)$, which is $8/9$.

(ii) Compute the following joint probability. Simplify your answer as much as possible and express it in terms of the parameters p_v, p_a, q_0, q_1 and r (you might not need all of them).

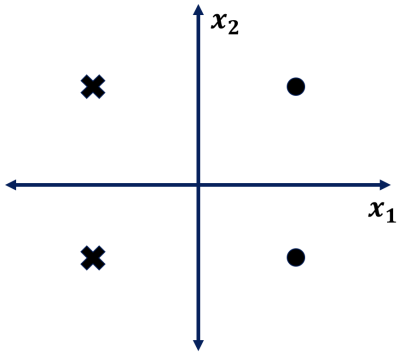
$$P(Y^{(20)}=1, X_v^{(20)}=1, X_a^{(20)}=1, Y^{(19)}=1, Y^{(18)}=1) = \underline{\quad p_a p_v [q_0^3(1-r) + q_1^3 r] \quad}$$

$$\begin{aligned}
&P(Y^{(20)}=1, X_v^{(20)}=1, X_a^{(20)}=1, Y^{(19)}=1, Y^{(18)}=1) \\
&= \sum_z P(Y^{(20)}=1|Z^{(2)}=z) P(Z^{(2)}=z) P(X_v^{(20)}=1|Y^{(20)}=1) P(X_a^{(20)}=1|Y^{(20)}=1) \\
&\qquad P(Y^{(19)}=1|Z^{(2)}=z) P(Y^{(18)}=1|Z^{(2)}=z) \\
&= q_0(1-r) p_a p_v q_0 q_0 + q_1 r p_a p_v q_1 q_1 \\
&= p_a p_v [q_0^3(1-r) + q_1^3 r]
\end{aligned}$$

Q2. Perceptrons and Naive Bayes

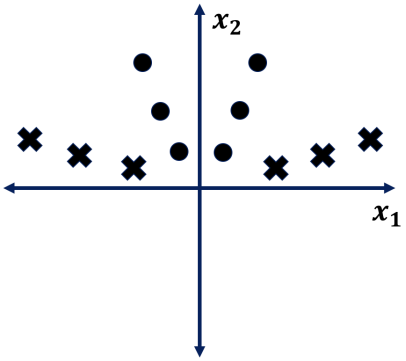
(a) For each of the datasets represented by the graphs below, please select the feature maps for which the perceptron algorithm can perfectly classify the data.

Each data point is in the form (x_1, x_2), and has some label Y, which is either a 1 (dot) or -1 (cross).



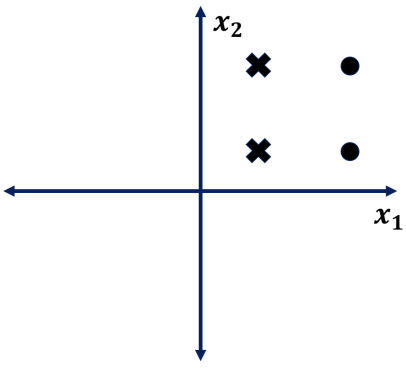
(i)

- [x] $[x_1 \quad x_2 \quad 1]$
- [x] $[x_1 \quad x_2 \quad x_1^2]$
- [x] $[x_1 \quad x_2 \quad |x_1|]$
- [x] $[x_1 \quad x_2 \quad Y]$
- [x] $[x_1 \quad x_2]$



(ii)

- [ ] $[x_1 \quad x_2 \quad 1]$
- [x] $[x_1 \quad x_2 \quad x_1^2]$
- [x] $[x_1 \quad x_2 \quad |x_1|]$
- [x] $[x_1 \quad x_2 \quad Y]$
- [ ] $[x_1 \quad x_2]$



(iii)

- [x] $[x_1 \quad x_2 \quad 1]$
- [x] $[x_1 \quad x_2 \quad x_1^2]$
- [ ] $[x_1 \quad x_2 \quad |x_1|]$
- [x] $[x_1 \quad x_2 \quad Y]$
- [ ] $[x_1 \quad x_2]$

(b) Performing maximum likelihood estimation (MLE) to fit the parameters of a Bayes net to some given data (with no Laplace smoothing) leads to which of the following learning algorithms?

- [x] Naive Bayes
- [ ] Perceptrons
- [ ] Kernelization
- [ ] Neural Networks
- ○ None

(c) Suppose that we are trying to perform a binary classification task using Naive Bayes. Y is the label, and (X_1, X_2) are the features. The domain for the features is anywhere on the 3×3 grid centered at $(0, 0)$. In other words, X_1 and X_2 have the domain $\{-1, 0, 1\}$

Suppose that this is your dataset: (0, 1, +), (0, −1, −), (−1, 1, +), (-1, −1, −), (1, 0, +), (−1, 1, −), (0, 0, +). What is the learned value of each of the following? (Leave your answer as a simplified fraction)

(i) $P(Y = +)$

$\frac{4}{7}$

(ii) $P(X_1 = 1|Y = -)$

0

(iii) $P(X_2 = 0|Y = +)$

$\frac{2}{4}$

(d) Now, to decouple from the previous question, assume that the learned CPTs are below.

Y	X_1	$\Pr(X_1 \mid Y)$
+	−1	0.4
+	0	0.1
+	1	0.5
−	−1	0.6
−	0	0.3
−	1	0.1

Y	X_2	$\Pr(X_2 \mid Y)$
+	−1	0.2
+	0	0.2
+	1	0.6
−	−1	0.7
−	0	0.1
−	1	0.2

Y	$\Pr(Y)$
+	0.2
−	0.8

(i) What would be the predicted value for Y if the data point is at $(0, 0)$?

$\hat{Y} = -$

$P(Y = -, x_1 = 0, x_2 = 0) = 0.8 * 0.3 * 0.1 = 0.024$, $P(Y = +, x_1 = 0, x_2 = 0) = 0.2 * 0.1 * 0.2 = 0.004$

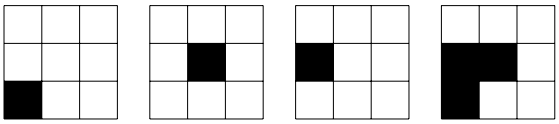
(ii) What would be the predicted value for Y if the data point is at $(1, -1)$?

$\hat{Y} = -$

$P(Y = -, x_1 = 1, x_2 = -1) = 0.8 * 0.1 * 0.7 = 0.056$, $P(Y = +, x_1 = 1, x_2 = -1) = 0.2 * 0.5 * 0.2 = 0.02$

Q3. [Timed: 15 Mins]ML: Perceptrons and Kernels

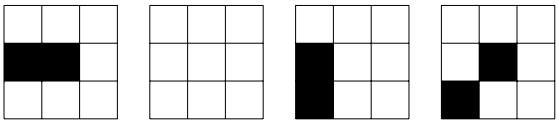
You've decided to single-handedly advance AI by constructing a perfect classifier to separate pictures of dogs and cats. With your state of the art 9-pixel camera, you've taken 4 pictures of dogs and 4 pictures of cats. These are the pictures of dogs (Class +):



(D-1) (D-2) (D-3) (D-4)

Class +

And these are the pictures of cats (Class -): (the cat is hiding in the second picture)



(C-1) (C-2) (C-3) (C-4)

Class -

You decide to mathematically model the dataset as:

x_1	x_2	x_3
x_4	x_5	x_6
x_7	x_8	x_9

$= (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9) = x$

where $x_i = 1$ if the corresponding pixel is black and 0 otherwise.

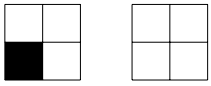
(a) Impressed with the quality of your photos, you want to run the perceptron algorithm with the weight vector initialized to $w_0 = (0, 0, 0, 1, 1, 1, 1, 1, 1)$. When your classifier gives a positive score, it predicts dog. When your classifier gives a negative score it predicts cat. To break ties, since you like cats more than dogs, when your classifier gives a score of 0, it predicts cat. Using the ordering left to right, dogs first and then cats:

(i) What is the first misclassified training datum? the first cat

(ii) What is the weight vector after the first perceptron update? It misclassifies the first cat, so the update is $w_5 = (0, 0, 0, 0, 0, 1, 1, 1, 1)$

(b) At heart, you're an artist. You feel the pictures of cats and dogs would look better if you changed them a little. Under which of the following transformations is the dataset above **NOT linearly separable**?

- The identity transformation, $\phi(x) = x$. In other words, the dataset above is not linearly separable.
- $\phi(x) = \bar{x}$. That is, if a cell is black it is turned white and vice versa.
- A rotation 90 degree clockwise
- A horizontal reflection
- The number of times each of the following patterns appear:
  (e.g. D-1 contains the first pattern once and the second pattern three times.)



(c) Consider the original feature space (i.e. none of the transformations above have been applied). Indicate whether a weight vectors can be constructed that makes the perceptron equivalent to the decision rules below. If the answer is yes, then include such a weight vector. Remember, when your classifier gives a positive score, it predicts dog. When your classifier gives a negative score it predicts cat. To break ties, since you like cats more than dogs, when your classifier gives a score of 0, it predicts cat.

Predict dog (+) if at least 3 squares are black, otherwise predict cat.

No

Predict dog (+) if any two adjacent squares are black, otherwise predict cat.

No

Predict dog (+) if any square other than a corner square is black, otherwise predict cat.

Yes $w = (0, 1, 0, 1, 1, 1, 0, 1, 0)$

Predict dog (+) if the photo is symmetric with respect to a 90 degree rotation, otherwise predict cat.

No

Predict dog (+) if the photo is symmetric with respect to its horizontal and vertical reflections, otherwise predict cat.

No