ICTS 4156: Introduction to ML

Razvan Bunescu

Lecture notes, January 27, 2021

1 Quiz, week 1

1.1 Quiz question 2

Let w = [1, 2, 3] and x = [-1, 2, -3]. Then their dot-product is $\mathbf{w}^T\mathbf{x} = 1*-1+2*2+3*-3 = -6$.

1.2 Quiz question 2

There are 2 bowls of fruits: one bowl contains 1 apple and 1 orange, while the other bowl contains 2 apples and 3 oranges. Patrick picks one fruit from each bowl. What is the probability that he ends up with 2 apples?

Implicit (default): each fruit in a bowl is equally likely to be picked. In bowl b_1 we have two fruits, a and o. Question is what is $P(a|b_1) = P(F = a|B = b_1)$.

There are two possible choices (outcomes) for (the random variable) F: either (mutually exclusive) a or o.

According to basic probability theory rules (axiom), $P(a|b_1) + P(o|b_1) = 1$. (1)

But, equally likely to be picked means $P(a|b_1) = P(o|b_1)$ (2)

From (1) and (2), it results that $P(a|b_1) = P(o|b_1) = 1/2$

$P(a_1|b_2) = 1/5$ and $P(a_2|b_2) = 1/5$. Then $P(a|b_2) = P(F = a_1 \vee F = a_2) = P(a_1|b_2) + P(a_2|b_2) = 1/5 + 1/5 = 2/5$

$P(A \vee B) = P(A) + P(B)$ whenever A and B are mutually exclusive events.

Probability that Patrick picks an apple a from bowl 1 (b_1) is $P(a|b_1) = 1/2$. Probability that Patrick picks an apple a from bowl 2 (b_2) is $P(a|b_2) = 2/5$.

What is the probability that he ends up with 2 apples? This is $P(E_1 = a \wedge E_2 = a) = P(a|b_1) \cdot P(a|b_2) = 1/2 \cdot 2/5 = 1/5 = 0.2$

$P(A \wedge B) = P(A) \cdot P(B)$ whenever A and B are (mutually) independent events.

1.3 Quiz question 3

There are 2 bowls of fruits: one bowl contains 1 apple and 1 orange, while the other bowl contains 2 apples and 3 oranges. Patrick picks one fruit from each bowl. What is the probability that he ends up with 1 apple and 1 orange (order does not matter)?

$P(a|b_1) = 1/2.$
$P(o|b_2) = 3/5.$
$P(a, o) = P(a|b_1) \cdot P(o|b_2) = 1/2 * 3/5$
$P(o, a) = P(o|b_1) \cdot P(a|b_2) = 1/2 * 2/5$
$P(a, o) + P(o, a) = 0.3 + 0.2 = 0.5$

1.4 Quiz question 4

Consider the following training dataset:

x1 = [1, 0, 1, 1], label y1 = +1

x2 = [1, 0, 0, 1], label y2 = +1

x3 = [1, 1, 1, 1], label y3 = -1

Which of the following parameters $\mathbf{w}$ perfectly fit the data, meaning $\mathbf{w}^T\mathbf{x} > 0$ if and only if $y_j = +1$, for all j = 1, 2, 3.

2 Linear algebra

Let $\mathbf{w} = [w_1, w_2, ..., w_K]$ and $\mathbf{x} = [x_1, x_2, ..., x_K]$ be two real-valued vectors. Equivalently, $\mathbf{w}, \mathbf{x} \in R_{K\times 1}$. Then, the dot-product of $\mathbf{w}$ and $\mathbf{x}$ (also called their inner-product) is calculated as:

$$\begin{aligned} \mathbf{w}^T\mathbf{x} &= \sum_{k=1}^{K} w_k * x_k && (1)\\ &= w_1 * x_1 + w_2 * x_2 + ... + w_K * x_K && (2)\\ &= w_1x_1 + w_2x_2 + ... + w_Kx_K && (3) \end{aligned}$$

By default, vectors are considered to be column vectors, i.e. $\mathbf{w}, \mathbf{x}$ have K rows and one column.

If we have two matrices, $A \in R_{M\times N}$ and $B \in R_{N\times K}$, then $C = A * B$, where $C \in R_{M\times K}$.

$$c_{i,j} = \sum_{n=1}^{N} a_{i,n} b_{n,j} \quad (4)$$

We can see the two vectors $\mathbf{w}, \mathbf{x}$ as two dimensional matrices, i.e. $\mathbf{w}, \mathbf{x}$ have K rows and one column. Then $\mathbf{w} \in R_{K\times 1}$ and $\mathbf{x} \in R_{K\times 1}$. However, $\mathbf{w}^T \in R_{1\times K}$ and $\mathbf{x} \in R_{K\times 1}$. Then, we can multiply them as inner-product $\mathbf{w}^T\mathbf{x}$ (which is a scalar) or as outer-product $\mathbf{x}\mathbf{w}^T$ (which is $K \times K$).

What do we mean by the (Euclidean) norm $||\mathbf{w}|| = ||\mathbf{w}||_2$ of a vector $\mathbf{w}$?

$$||\mathbf{w}|| = \sqrt{\sum_{k=1}^{K} w_k^2} \tag{5}$$

$$||\mathbf{w}||^2 = \sum_{k=1}^{K} w_k^2 \tag{6}$$

3 Probability theory

Two colors: *red* and *blue*. We designed one feature, $\phi_1(\mathbf{x}) = 1$ if and only if the object is *blue*, i.e. if $\mathbf{x}$ is *red*, $\phi_1(\mathbf{x}) = 0$.

Q: Can we use $\phi_1(\mathbf{x}) = 2$ instead of $\phi_1(\mathbf{x}) = 1$?

Will use x_1 instead of $\phi_1(\mathbf{x})$.

The only way x_1 is used when doing classification with our simple linear model is through the product $w_1 x_1$.

Replacing $x_1 = 1$ with $x_1 = 2$ results in w_1 being replaced in the overall dot-product with $2w_1$.

However, $1 * w_1 = 2 * (w_1/2) = 2 * w_1'$, where $w_1' = w_1/2$.

4 Notes on lecture slides material

Suppose we have three colors: *red* and *blue* and *green*. How should we encode this property as feature(s)?

1. Use 3 Boolean features: $x_1 = 1$ iff x is red, $x_2 = 1$ iff x is blue, and $x_3 = 1$ iff $\mathbf{x}$ is green.

2. Use 1 real-valued feature $x_1 = 0$ if x is red, $x_1 = 1$ if x is blue, and $x_1 = 2$ if $\mathbf{x}$ is green.

   - This is used as $w_1 x_1$ when computing the "score" $\mathbf{w}^T\mathbf{x}$ for object $\mathbf{x}$. By using the values above, we tell the ML model, before it even gets a chance to learn the parameters $\mathbf{w}$, that the color red does not matter. We also tell it that being green ($w_1 * 2$) is twice as 'important' than being blue ($w_1 * 1$). **But there is no natural, relevant for our classification task, ordering between the colors**.

Suppose I want to train a ML model to predict whether my neighbor will wear a t-shirt, based on information such as air *temperature* T, whether the neighbor is at home, ...

Let's use one feature $x_1 = T$. Think about two situations: one where $x_1 = 10$ and one where $x_1 = 80$.