# Week 4: Naive Bayes, Entropy and Backpropagation

http://mlvu.github.io

March 3, 2021

## 1 preliminaries: probability

Before we start, it's helpful to review your general grasp of probability. If the exercises below give you any trouble we recommend that you follow the links provided to brush up a little.

**question 1:**

1. In a few sentences, explain the difference between the Frequentist and the Bayesian interpretation of probability.
2. What is the difference between a sample space and an event space?
3. What is the difference between a probability distribution and a probability density function?

**question 2:** In the following, $p$ is a probability function and $A$ and $B$ are random variables. Which of the following are true?

1. Joint probability is symmetric: $p(A, B) = p(B, A)$.
2. Conditional probability is symmetric $p(A \mid B) = p(B \mid A)$.
3. Two random variables $X$ and $Y$ are conditionally independent on a third $Z$. Once we know $X$ and $Y$, we also know the value of $Z$.
4. Two random variables $X$ and $Y$ are conditionally independent on a third $Z$. Once we know $Z$, also knowing $X$ will tell us nothing extra about $Y$.

**question 3:** Assume that the probability that a given patient has diabetes is $0.01$. We have a test for diabetes with a false positive rate of $0.05$: if a patient has no diabetes, the test diagnoses it 5% of the time. The false negative rate is $0.1$. You are a doctor, and you administer the test to a patient (knowing nothing else). The test says she has diabetes. What is the probability that she doesn't? Hint: this is a question about Bayes' rule. Reflect on the result. Is this what you would've expected? If not, where does the unexpected result come from?

More practice? Follow these links:

1. https://seeing-theory.brown.edu/ (especially this one)
2. https://betterexplained.com/articles/a-brief-introduction-to-probability-statistics/
3. https://www.khanacademy.org/math/probability/probability-geometry/probability-basics/a/probability-the-basics
4. http://dept.stat.lsa.umich.edu/~moulib/probrefresh.pdf

## 2 Naive Bayes

The following dataset represents a spam classification problem: we observe 8 emails and measure two binary features with the values True and False. The first is True if the word "pill" occurs in the e-mail and False if it doesn't, and the second is True if the word "meeting" occurs and False if it doesn't.

| "pill" | "meeting" | label |
|---|---|---|
| T | F | Spam |
| T | F | Spam |
| F | T | Spam |
| T | F | Spam |
| F | F | Ham |
| F | F | Ham |
| F | T | Ham |
| T | T | Ham |

**question 4:** We will build a naive Bayes classifier for this data. What is the defining property of naive Bayes? Why is it called "naive"?

**question 5:** We build a naive Bayes classifier on this data, as described in the lecture. We get an email that contains both words. Which class does the classifier assign?

**question 6:** Which probabilities does the classifier assign to each class?

To improve our accuracy, we add another feature:

| "pill" | "meeting" | "hello" | label |
|---|---|---|---|
| T | F | F | Spam |
| T | F | F | Spam |
| F | T | F | Spam |
| T | F | F | Spam |
| F | F | F | Ham |
| F | F | F | Ham |
| F | T | F | Ham |
| T | T | T | Ham |

**question 7:** For the class Spam, there are no emails recorded that contain the word "hello". Why is this a problem?

What solution is suggested in the slides?

**question 8:** Implement this solution, and give the class probabilities for an email containing all three words.

## 3 Entropy

We define two probability distributions $p$ and $q$ on a set of four outcomes $\{a, b, c, d\}$.

| $p(a)$ | $p(b)$ | $p(c)$ | $p(d)$ |
|---|---|---|---|
| $1/4$ | $1/4$ | $1/4$ | $1/4$ |

| $q(a)$ | $q(b)$ | $q(c)$ | $q(d)$ |
|---|---|---|---|
| $1/2$ | $1/4$ | $1/8$ | $1/8$ |

**question 9:** Could you simulate sampling from these distributions using coinflips as described in the lecture?

**question 10:** Compute the entropy of $p$ and $q$.

The entropy of $q$ is lower than the entropy of $p$. What does this tell you about the difference between the two distributions?

**question 11:** In the definition of entropy that we use (information entropy), the logarithms have base 2 (i.e. $\log_2$ instead of $\log_{10}$ or $\ln$). This follows directly from our decision to model probability distributions with coinflips. How?

## 4 Backpropagation

We will practice the backpropagation algorithm as described in the slides. We will use a neural network defined by the following function:

$$\begin{aligned} y &= v_1 h_1 + v_2 h2 \\ h_1 &= \sigma(k_1) \\ h_2 &= \sigma(k_2) \\ k_1 &= w_1 x \\ k_2 &= w_2 x \end{aligned}$$

Where $\sigma$ represents the logistic sigmoid. The network has a single input node $x$ and a single output node $y$. The weights are $w_1$, $w_2$, $v_1$ and $v_2$. We've left out bias nodes to keep things simple.

**question 12:** Draw this network.

We will use stochastic gradient descent to train this network. That means we define the loss function for a single instance. We will use basic least-squares loss to train this network for regression. Our loss function for instance $x$ is

$$\text{loss}_x(w_1, w_2, v_1, v_2) = \frac{1}{2}(y - t)^2$$

where $t$ is the target value provided by the dataset, and $y$ is the output of the network.

We see each line in the definition above as a module, with some inputs and some outputs. Using the chain rule, we will express the derivative with respect to weight $w_1$. We will use only *local derivatives*, expressing the derivative for each module with respect to its inputs, but not working out the derivative beyond that.

**question 13:** Fill in the gaps:

$$\begin{aligned}\frac{\partial \text{loss}}{\partial w_1} &= \frac{\partial \text{loss}}{\partial y}\frac{\partial y}{\partial h_1}\cdots\frac{\partial k_1}{\partial w_1} \\ &= (y - t) \times \ldots \times \sigma(k_1)(1 - \sigma(k_1)) \times \ldots \end{aligned} \tag{1}$$

**question 14:** The network has a diamond shape, just as shown in the slide when explaining the multivariate chain rule (in the *Deep Learning 1* lecture). However, in this case, we don't need the multivariate chain rule. Why not? In what kind of situation would the multivariate chain rule be required?

We could take the formulation above and fill in the symbolic expressions for $y$, $k_1$, etc, and come up with a general symbolic formula for the derivative for all inputs. However, that is usually expensive to do for large neural nets. Instead, we leave it as is, and fill in the *numeric* values for these variables for a specific input and for specific weights.

**question 15:** Assume that the input is $x = 1$, with target output $t = \frac{1}{2}$ and that all weights are set at 1. Do a *forward pass*: compute the loss and all intermediate values $k_1$, $k_2$, $h_1$, $h_2$ and $y$.

To simplify calculation, you can use the approximation $\sigma(1) = \frac{3}{4}$.

Now we do the *backward pass*. Fill these intermediate values in to the loss function decomposed by the chain rule from line 1, and compute the derivative with respect to $w_1$ for these inputs and weights.