multinomial_distro

October 12, 2020

[11]:
```
import random

# generative models can generate data. Here is a very simple data generating␣
↪function with one Bernoulli parameter P
def generatePoint(p):

    if (random.random() < p):
        return 1
    else:
        return 0

p = .5

data = []

for i in range(10):
    data.append(generatePoint(0.5))

data
```

[11]: [0, 0, 1, 0, 0, 0, 1, 0, 0, 0]

0.0.1 Question

If you vary p what do you observe about your generated data?

p closer to 1 results in more ones

0.0.2 Learning

Usually you don't know the parameters of your distribution when you start. You just have data. Let's generate some data with an "unknown" p, meaning it is "hidden" in a Python variable. We just won't peek at the variable. This is what things are like in the real world. You just have data; you don't know how it was generated.

[11]:
```
secret_p_from_nature = random.random()

data_in_the_world = []
```

```
for i in range(10):
    data_in_the_world.append(generatePoint(secret_p_from_nature))
```

[12]:
```
data_in_the_world
```

[12]: `[0, 0, 1, 1, 1, 0, 1, 1, 1, 0]`

0.0.3 Question

What do you think `secret_p_from_nature` is? Try to guess the value based on `data_in_the_world`. But don't peek!!

0.6

0.0.4 Learning 2

How can we systematically decide what parameter do we think generated the data? To answer we again use the "likelihood function". You have already seen the likelihood function when we studied logistic regression. The likelihood function returns the probability of the data X given the parameters θ.

We write the likelihood function as $\mathcal{L}(\theta|X) = p_\theta(X)$ where θ is your parameters.

- You should read this as: if the parameters are θ what is the probability of our data?
- Here $\theta = p$.
- Notice that the likelihood function is a function of θ.

0.0.5 Question

The likelihood function is a function. Functions in general map inputs to outputs.

- What is the input to the function?
- What is the output?

The input is θ and the output is the probability that we see the data X with the given parameters

0.0.6 Defining the likelihood

So far we have talked about the likelihood as an abstraction. But what is $p_\theta(X)$?

This is something that *we* specify as the practitioner. When you make a generative model you are observing some data from nature and claiming that the data was generated in a particular way. You are making a claim about the process that generated the data. That process has parameters θ. The process + the parameters give you the likelihood.

0.0.7 Defining the likelihood 2

For our case, let's say the data was generated from a Bernoulli distribution. A Bernoulli distribution defines the probability of a binary event, which can be thought of as a yes or no. For instance, you might use a Bernoulli distribution to model the probability that yes a coin lands on heads. The Bernoulli distribution has one parameter p.

According to the Bernoulli distribution, the probability of a datapoint x_i given p is

$$f(x_i; p) = \begin{cases} p & \text{if } x_i = 1 \\ q = 1 - p & \text{if } x_i = 0 \end{cases} \tag{1}$$

[16]:
```
def BernoulliProbOnePoint(p, x_i):
    '''
    return the probability of x_i, according to the Bernoulli distribution with␣
 ↪parameter p
    '''

    if(x_i):
        return p
    else:
        return 1 - p
```