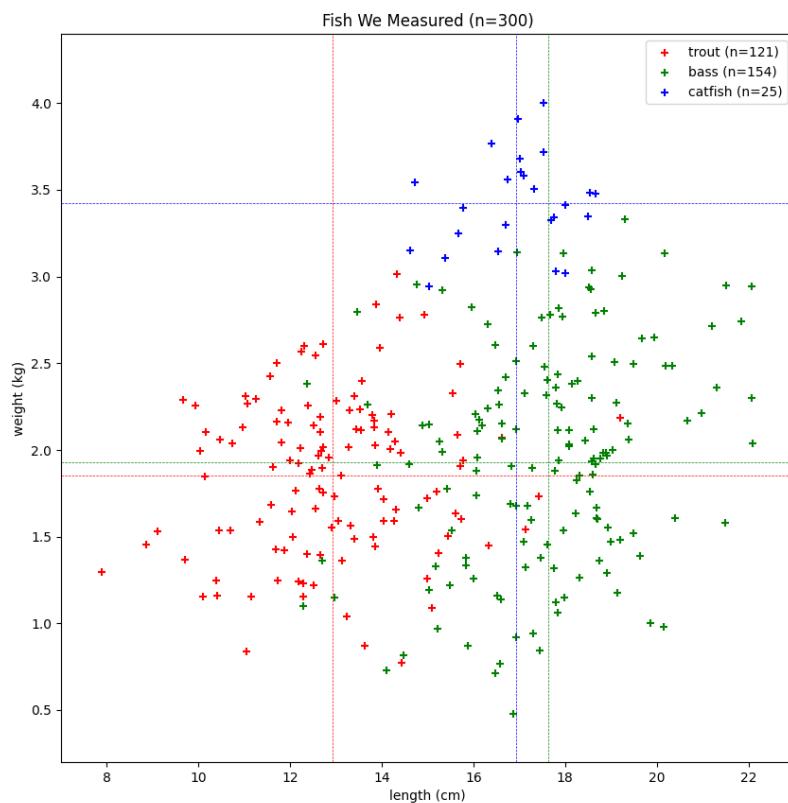


CHAPTER 1

Bayesian Classifiers

You drop a net in several places in a lake. You measure, weigh, and identify every fish to be one of the following: trout, bass, or catfish. Here is a scatter plot of what you find:



(Data: `make_data.py`, Graph: `1_scatter_fish.py`)

The vertical and horizontal lines represent the mean length and width (respectively) of each type of fish.

You are writing a system that someone on the lake could use to identify their fish.

1.1 The Prior

Even before the person measures and weighs the fish, they can make a decent guess at what any randomly selected fish is:

Fish	Count	Percent
trout	121/300	40.3%
bass	154/300	51.3%
catfish	25/300	8.3%

If you know nothing about the fish, there is a 51.3% chance it is a bass. The probability without any observations is known as the *prior*:

$$P(\text{fish} = \text{trout}) = .403$$

$$P(\text{fish} = \text{bass}) = .513$$

$$P(\text{fish} = \text{trout}) = .083$$

However, now the user makes an observation: The fish is 16 inches long and weighs 2.1 kg. Can we update our beliefs based on this? Sounds like a job for conditional probability.

By the definition of conditional probability:

$$P(\text{fish} = \text{trout} | \text{length} = 16, \text{weight} = 2.1) = \frac{p(\text{fish} = \text{trout}, \text{length} = 16, \text{weight} = 2.1)}{p(\text{length} = 16, \text{weight} = 2.1)}$$

At this rate, we are going to run out of space, so we will use F to represent fish, L to represent length, and W to represent weight.

We can calculate the denominator by summing up all the possibilities. Now, the right-hand side is:

$$\frac{p(F=\text{trout}, L=16, W=2.1)}{p(F=\text{trout}, L=16, W=2.1) + p(F=\text{bass}, L=16, W=2.1) + p(F=\text{catfish}, L=16, W=2.1)}$$

We can calculate the probability of all three possibilities (which will add up to 1.0).

1.2 Naive Bayes

The assumption in naive Bayes is that the probability distributions of length and weight are conditionally independent given the breed of the fish. That is:

$$p(F=\text{trout}, L=16, W=2.1) = p(L=16|F=\text{trout})p(W=2.1|F=\text{trout})P(F=\text{trout})$$

Within a particular species, things like length and weight, which represent of the sum of a lot of genetic and environmental factors, are often normally distributed.

What if we assume that length has a normal distribution for each kind of fish? What is the most likely estimator for the mean of each trait?

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

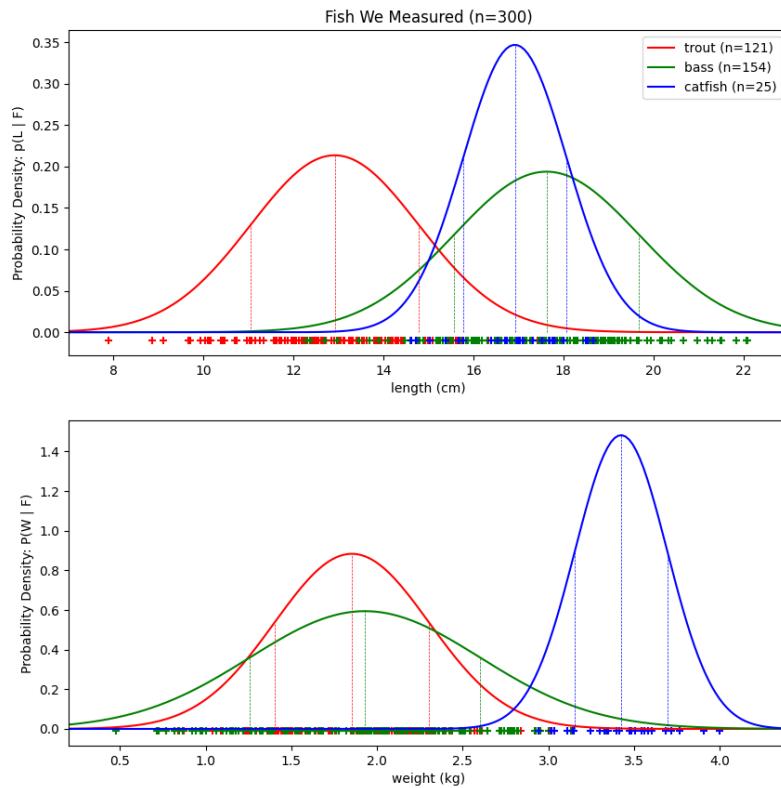
And variance?

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (\mu - x_i)^2$$

We can compute those pretty easily:

Fish	Trait	μ	σ^2
Trout	Length	12.91522571	3.49054363
	Weight	1.85290281	0.20387574
Bass	Length	17.62628327	4.23881747
	Weight	1.92960088	0.45092986
Catfish	Length	16.92187387	1.32301935
	Weight	3.42423185	0.07244546

If we plot those probability distributions out:



(Source:2_univariates.py)

Note that the area under each one of those curves is exactly 1.0. They are probability distributions.

Now we can use the formula for the normal distribution:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Thus,

Given	$p(L = 16 F)$	$p(W = 2.1 F)$
$F=\text{Trout}$.0546	.7607
$F=\text{Bass}$.1418	.5753
$F=\text{Catfish}$.2516	0.01

Using the naive assumption, we can now compute the joint probability for any type of fish:

$$p(F = ?, L = 16, W = 2.1) = p(W = 2.1|F = ?)p(L = 16|F = ?)p(F = ?)$$

For	$p(L = 16, W = 2.1, F = ?)$
F=Trout	.016763
F=Bass	.041886
F=Catfish	0.00001

The sum of these is $p(L = 16, W = 2.1)$: .0586.

By the definition of conditional probability:

$$p(F = ? | L = 16, W = 2.1) = \frac{p(F = ?, L = 16, W = 2.1)}{p(L = 16, W = 2.1)}$$

So,

For	$p(F = ? L = 16, W = 2.1)$
F=Trout	.286
F=Bass	.714
F=Catfish	≈ 0

If you pull a random fish out of the lake and its length is 16 and its weight is 2.1, there is a 71.4% chance that it is bass. There is a 28.6% chance it is a trout.

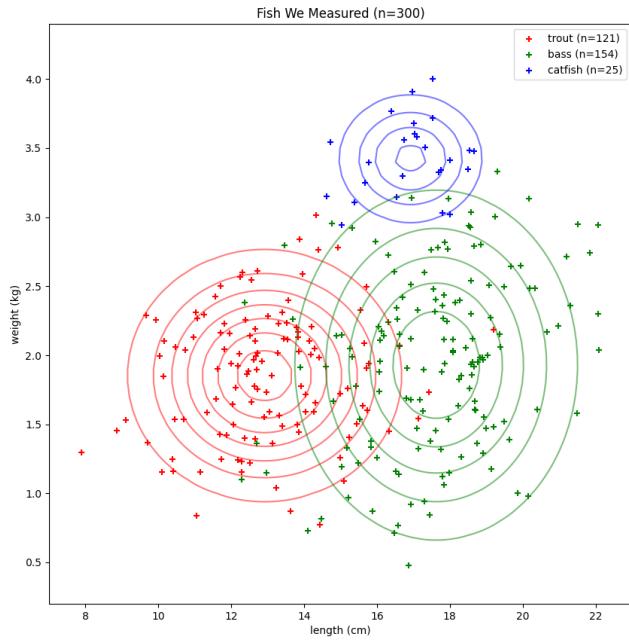
1.3 Using the multivariate Gaussian distribution

Given the kind of fish, length and width are independent!? That makes no sense. Longer fish tend to be heavier, right?

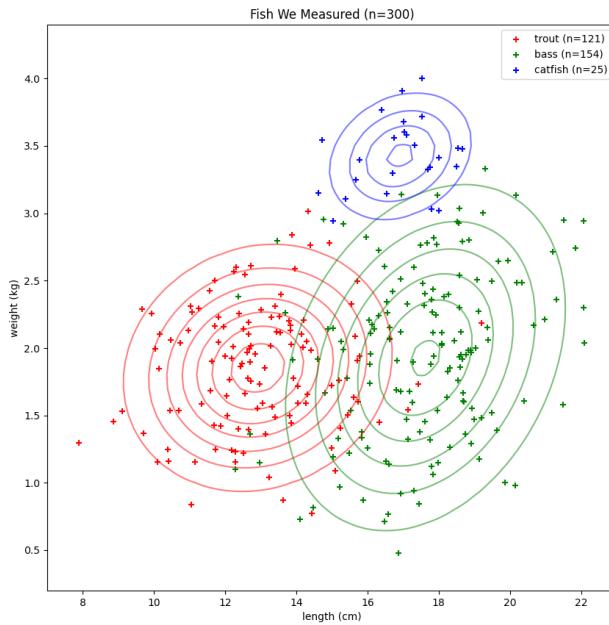
The multivariate Gaussian distribution is a generalization of the normal distribution::

- It deals with data of any dimension d.
- It can include the idea of covariance. For example, length and width are not independent.

If we use your naive Bayes, we can get two-dimensional density plots that would look like this:



With independence, there are ovals that are circles, ovals that go up, or ovals that sideways. If we use the multivariate Gaussian distribution, we get:



When there is covariance, the ovals can slant in any direction.

If we discard the naive assumption, we should get slightly more accurate results.

The multivariate Gaussian distribution uses vectors and matrices. The mean $\vec{\mu}$ is a vector

of dimension d . The variance of each variable and its covariance with the other variables is captured in a $d \times d$ matrix called *the covariance matrix* is usually called Σ . (Yes, this is also the symbol for summation. This creates some confusion at times, but you can usually figure out which is intended by the context.)

When you know $\vec{\mu}$ and Σ , the probability density for any vector \vec{x} is:

$$p(\vec{x}) = (2\pi)^{-d/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})\right)$$

If you have several data points $\vec{x}_1, \dots, \vec{x}_n$, how do you compute the values of $\vec{\mu}$ and Σ for which the observations $\vec{x}_1, \dots, \vec{x}_n$ would be most likely?

$$\vec{\mu} = \frac{1}{n} \sum_{i=1}^d \vec{x}_i$$

And the covariance matrix? (Here, we think of each \vec{x}_i and $\vec{\mu}$ as column vectors.)

$$\Sigma = \frac{1}{n} \sum_{i=1}^n (\vec{x}_i - \vec{\mu})(\vec{x}_i - \vec{\mu})^T$$

Notice that the diagonal entries of the matrix are the variance of each variable. Notice also that this is a symmetric matrix.

Applying this to our data we get:

Fish	$\vec{\mu}$	Σ
Trout	$\begin{bmatrix} 12.91522571 \\ 1.85290281 \end{bmatrix}$	$\begin{bmatrix} 3.51963149 & 0.09975744 \\ 0.09975744 & 0.20557471 \end{bmatrix}$
Bass	$\begin{bmatrix} 17.62628327 \\ 1.92960088 \end{bmatrix}$	$\begin{bmatrix} 4.26652216 & 0.39553268 \\ 0.39553268 & 0.45387711 \end{bmatrix}$
Catfish	$\begin{bmatrix} 16.92187387 \\ 3.42423185 \end{bmatrix}$	$\begin{bmatrix} 1.37814515 & 0.07281453 \\ 0.07281453 & 0.07546403 \end{bmatrix}$

We can use these to compute the likelihood of seeing a fish with length = 16 and weight = 2.1 for each type of fish:

Given	$p(L = 16, W = 2.1 F = ?)$
F=Trout	.04578
F=Bass	.07732
F=Catfish	.000001

Multiplied by the prior to get the joint probability:

Given	$p(L = 16, W = 2.1, F = ?)$
F=Trout	.01847
F=Bass	.03969
F=Catfish	≈ 0

We sum those to get $p(L = 16, W = 2.1) = .0582$.

Given	$p(F = ? L = 16, W = 2.1)$
F=Trout	.318
F=Bass	.682
F=Catfish	≈ 0

Given a randomly selected fish that is 19 inches long and 2.1 kg in weight, you would guess that it is a bass with a confidence of 68.2%.

This is a draft chapter from the Kontinua Project. Please see our website (<https://kontinua.org/>) for more details.

APPENDIX A

Answers to Exercises

