Exploring parameter space

Practice Quiz, 7 questions

7/7 points (100%)

Congratulations! You passed!

Next Item

1 / 1
point

1.

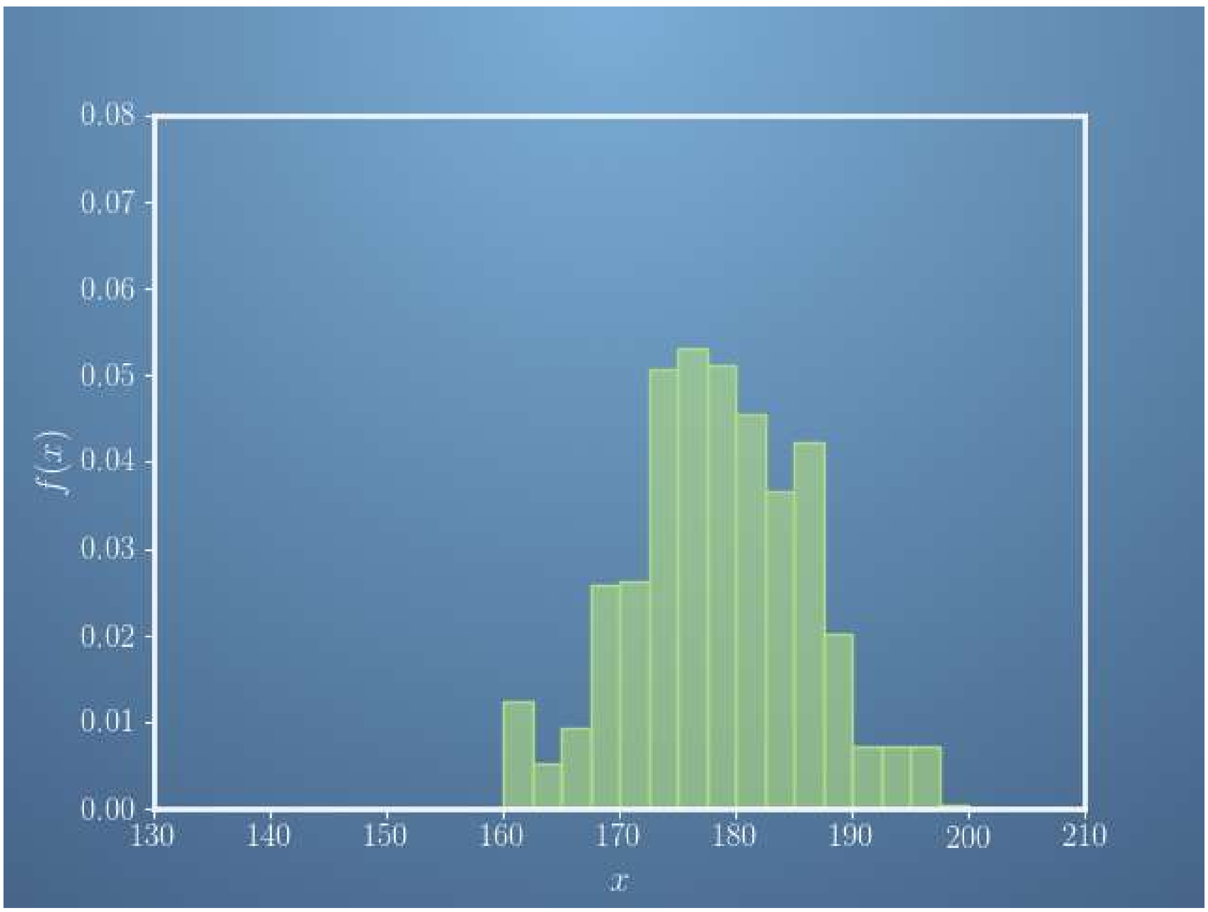
In this quiz, we shall see how quantities in machine learning can be represented as vectors. These could be in the form of data itself, or model parameters, and so on. We will begin to explore why framing our problem in this form might be a useful thing to do.

The problem we shall focus on in this exercise is the distribution of heights in a population.

Since a vector is just a list of numbers, one of the vectors that we can define relates to data that we measure. That is, in this case, we can record the frequency of people with heights between 150cm and 152.5cm, between 152.5cm and 155cm, and so on. We can define this as the vector $\mathbf{f}$ with components,

$$\mathbf{f} = \begin{bmatrix} f_{150.0-152.5} \\ f_{152.5-155.0} \\ f_{155.0-157.5} \\ f_{157.5-160.0} \\ f_{160.0-162.5} \\ \vdots \end{bmatrix}$$

This vector can also be represented by the histogram,



Of the following statements, select all that are true.

- [ ] If another sample was taken under the same conditions, the frequencies would be exactly the same.

Un-selected is correct

There are at least 1000 elements in the frequency vector, $\mathbf{f}$

Un-selected is correct

The sum of all the frequencies of the bins is one.

Un-selected is correct

There are at least 15 elements in the frequency vector, $\mathbf{f}$.

Correct
The data has been grouped into around 20 bins each having a width of 2.5cm

The total area of all the bins sums to one.

Correct
The area of a bin, which is its range times its frequency gives the probability that a person picked from the sample will have a height in that range.

None of the other statements.

Un-selected is correct

If another sample was taken under the same conditions, the frequencies would be broadly similar.

Correct
For a sufficiently large sample, the data will represent the population it is taken from.

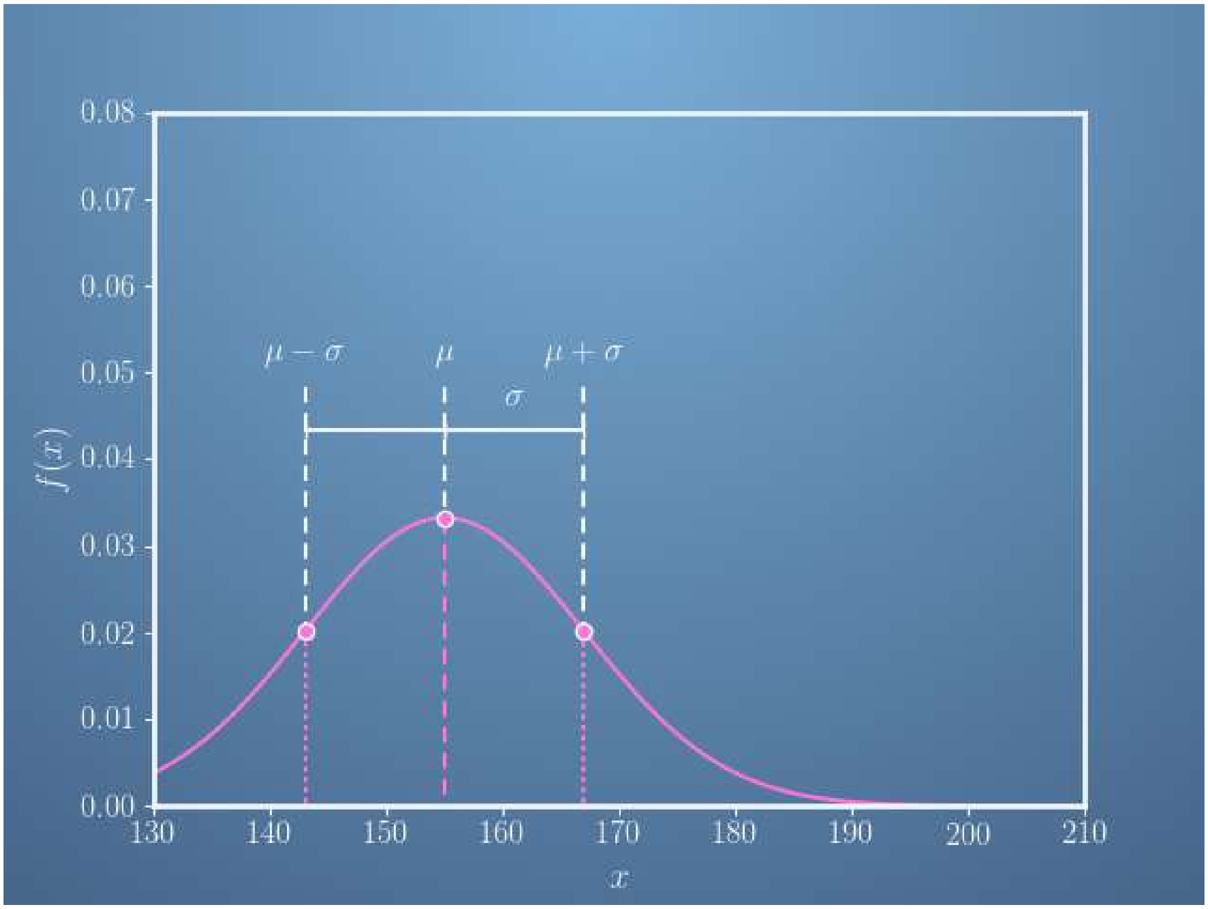
---

1 / 1
point

2.

One of the tasks of machine learning is to fit a model to data in order to predict the underlying distribution.

For the heights of a population, a model we may use to predict frequencies is the Normal (or Gaussian) distribution. This is a model for a bell-shaped curve, which looks like this,



It has an equation,

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

the exact form of which is unimportant, except that it is dependant on two parameters, the *mean*, μ, where the curve is centred, and the *standard deviation*, σ, which is the characteristic width of the bell curve.

We can put these two parameters in a vector, $\mathbf{p} = \begin{bmatrix} \mu \\ \sigma \end{bmatrix}$.

Pick the parameter vector $\mathbf{p}$ which best describes the distribution pictured.

$$\mathbf{p} = \begin{bmatrix} 155 \\ 24 \end{bmatrix}$$

$$\mathbf{p} = \begin{bmatrix} 167 \\ 24 \end{bmatrix}$$

$\mathbf{p} = \begin{bmatrix} 155 \\ 12 \end{bmatrix}$

Correct
The mean is 155cm and the standard deviation is 10cm.

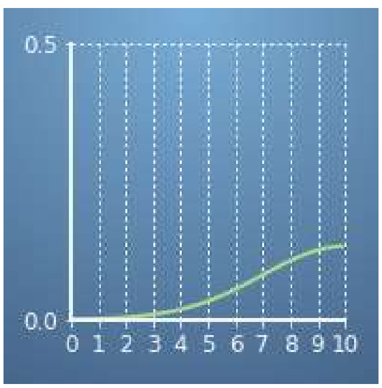
$\mathbf{p} = \begin{bmatrix} 167 \\ 12 \end{bmatrix}$

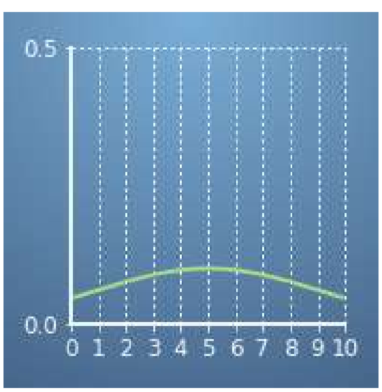
---

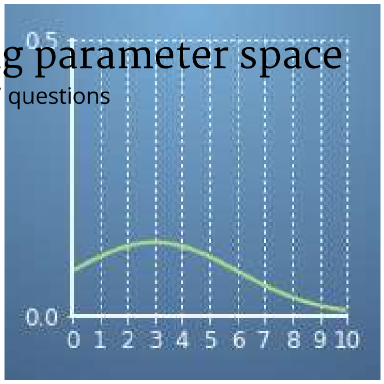
1 / 1
point

3.

Pick the Normal distribution that corresponds the closest to the parameter vector $\mathbf{p} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$.

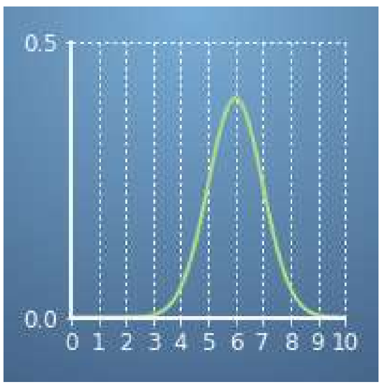






Correct

This distribution has parameters, $\mathbf{p} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$.



1 / 1
point

4.

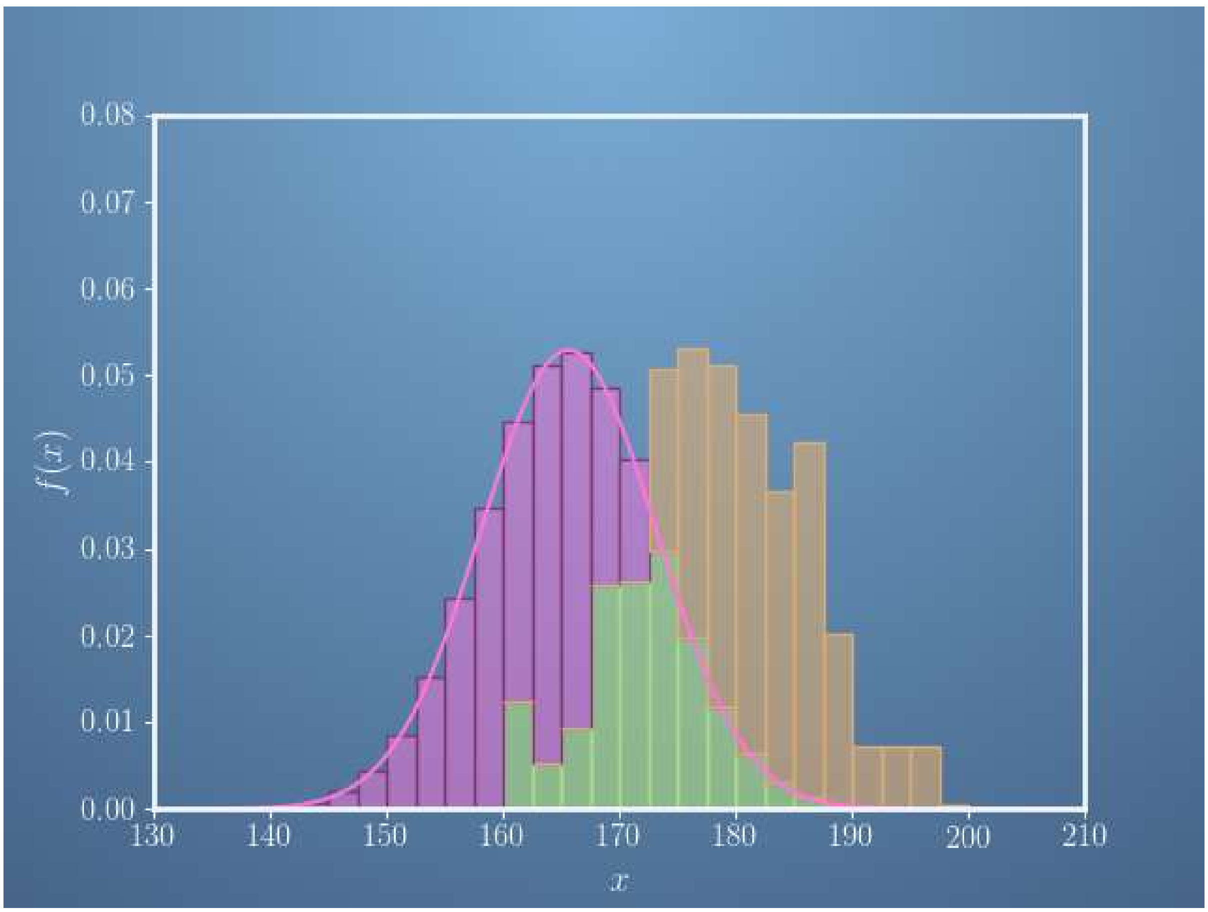
A model allows us to predict the data in a distribution. In our example we can start with a parameter vector $\mathbf{p}$ and convert it to a vector of expected frequencies $\mathbf{g_p}$, for example,

$$\mathbf{g_p} = \begin{bmatrix} g_{150.0-152.5} \\ g_{152.5-155.0} \\ g_{155.0-157.5} \\ g_{157.5-160.0} \\ g_{160.0-162.5} \\ \vdots \end{bmatrix}$$

A model is only considered good if it fits the measured data well. Some specific values for the parameters will be better than others for a model. We need a way fit a model's parameters to data and quantify how good that fit is.

One way of doing so is to calculate the "residuals", which is the difference between the measured data and the modelled prediction for each histogram bin.

This is illustrated below. The model is shown in pink, the measured data is shown in orange and where they overlap is shown in green. The height of the pink and orange bars are the residuals.



A better fit would have as much overlap as it can, reducing the residuals as much as possible.

How could the model be improved to give the best fit to the data?

- [ ] Increase the standard deviation, σ.

☐ Decrease the mean, μ.

Un-selected is correct

☒ Increase the mean, μ.

Correct
The mean of the model is too low.

☐ Decrease the standard deviation, σ.

Un-selected is correct

☐ Keep the mean, μ, approximately the same.

Un-selected is correct

☒ Keep the standard deviation, σ, approximately the same.

Correct
The model has a width similar to the data.

---

1 / 1
point

5.

The performance of a model can be quantified in a single number. One measure we can use is the *Sum of Squared Residuals*, SSR. Here we take all of the residuals (the difference between the measured and predicted data), square them and add them together.

In the language of vectors we can write this as, $\mathrm{SSR}(\mathbf{p}) = |\mathbf{f} - \mathbf{g}_\mathbf{p}|^2$, which will be explained further on in this course.

Use the following code block to play with parameters of a model, and try to get the best fit to the data.

```
# Play with values of μ and σ to find the best fit.
μ = 179 ; σ = 7.5
p = [μ, σ]
histogram(p)
```

Run

Reset

Find a set of parameters with a fit $\mathrm{SSR} \leq 0.00051$

Input your fitted parameters into the code block below.

```
# Replace μ and σ with values that minimise the SSR.
p = [179, 7.5]
```

Run

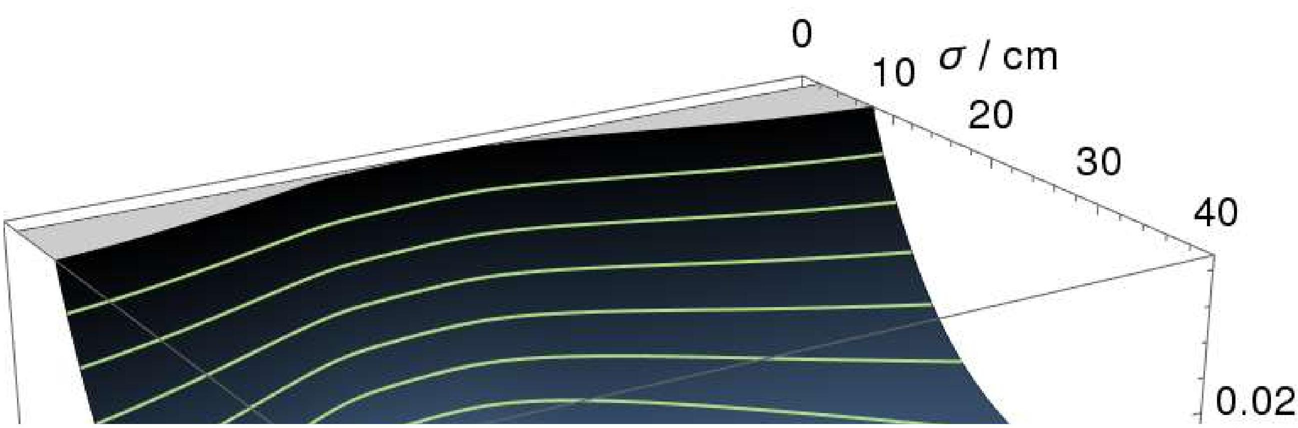
Reset

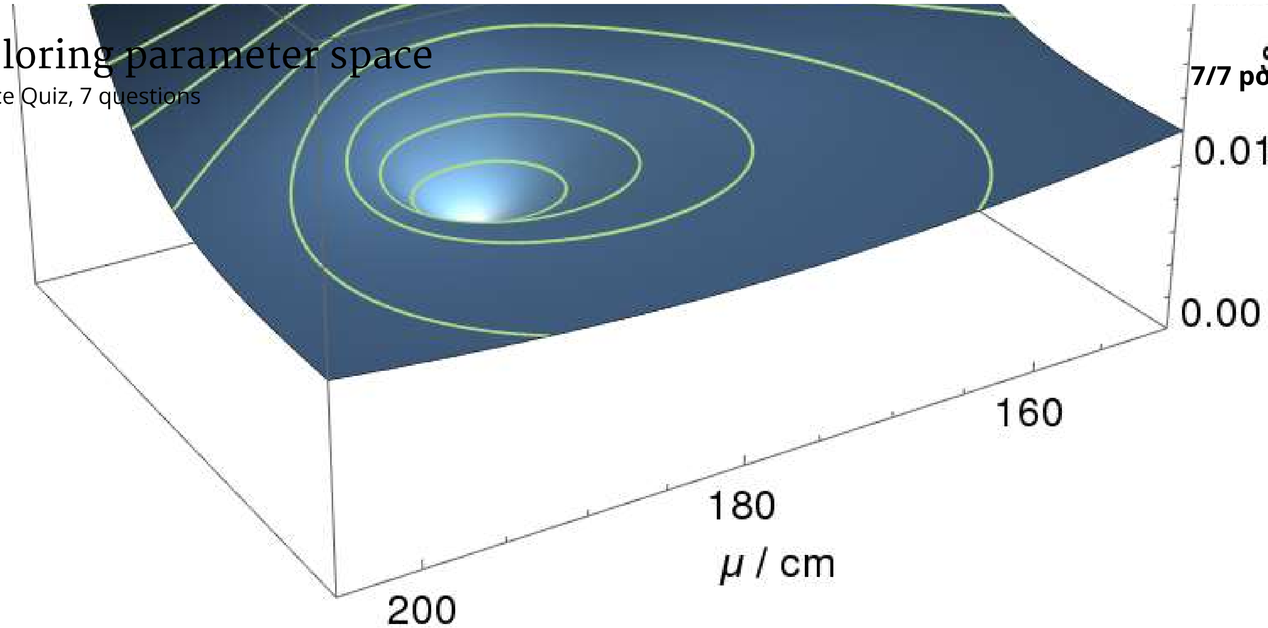
Correct Response
Well done! You found a model that fits the data acceptably well according the the criterion defined for SSR.

1 / 1
point

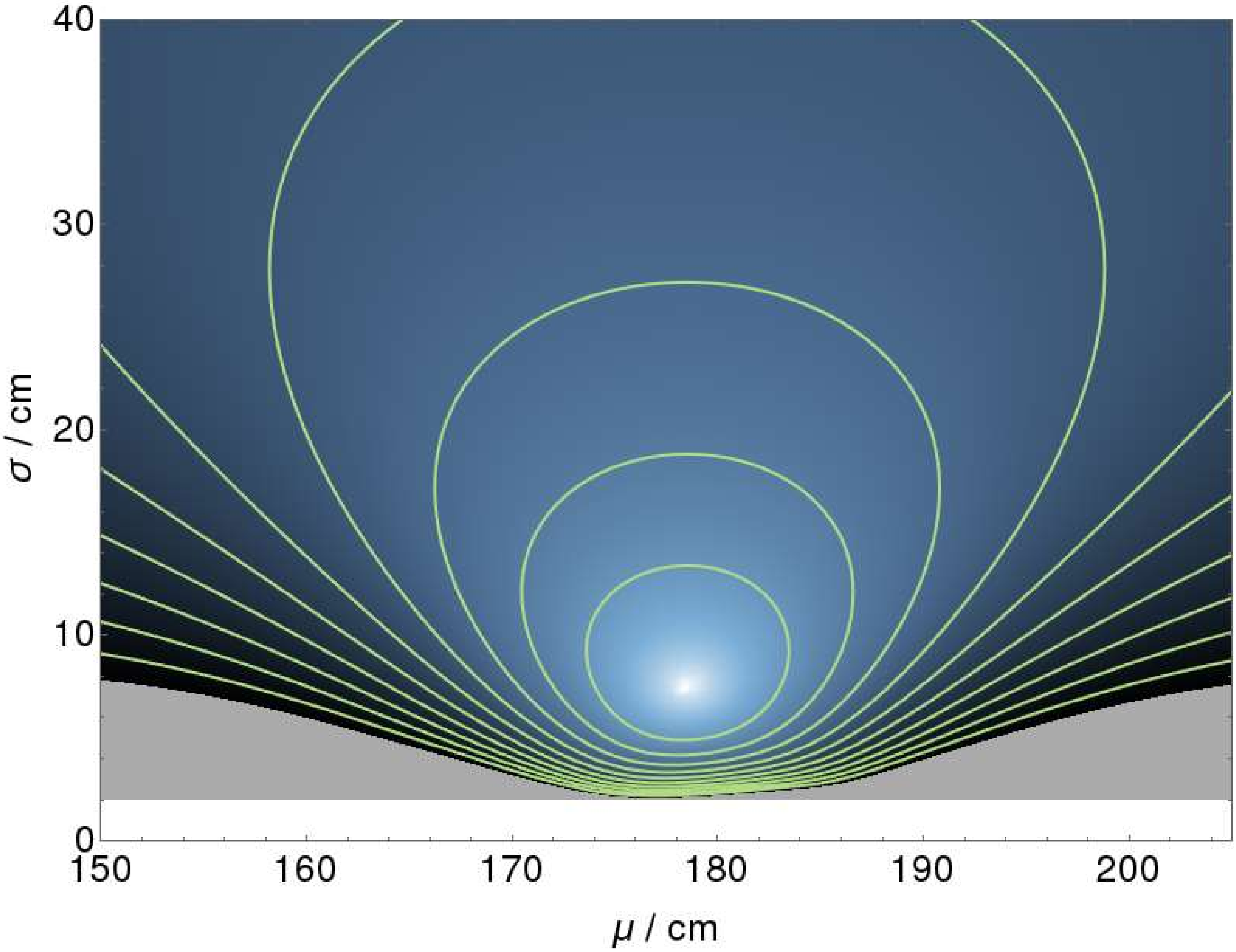
6.
Since each parameter vector $\mathbf{p}$ represents a different function, each with its own value for the sum of squared residuals, SSR, we can draw the surface of SSR values over the space spanned by $\mathbf{p}$, such as μ and σ in this example.

Here is an illustration of this surface for our data.





We can also take a 'top-down' view of the surface, and view it as a contour map, where each of the contours (in green here) represent a constant value for the SSR.



The goal in machine learning is to find the parameter set where the model fits the data as well as it possibly can. This translates into finding the lowest point, the global minimum, in this space.

Select all true statements below.

7/7 points (100%)

- [ ] None of the other statements.

Un-selected is correct

- [x] Each point on the surface represents a set of parameters $\mathbf{p} = \begin{bmatrix} \mu \\ \sigma \end{bmatrix}$.

Correct
This means each point in the space will generate a different histogram of expected data, which will perform better or worse against the measured data.

- [ ] You get the same model by following along a contour line.

Un-selected is correct

- [ ] At the minimum of the surface, the model exactly matches the measured data.

Un-selected is correct

- [x] Moving at right angles to contour lines in the parameter space will have the greatest effect on the fit than moving in other directions.

Correct
For example, moving along contour lines has no affect on the SSR (by definition). However moving perpendicular to them can significantly improve or reduce the quality of the fit.

---

1 / 1
point

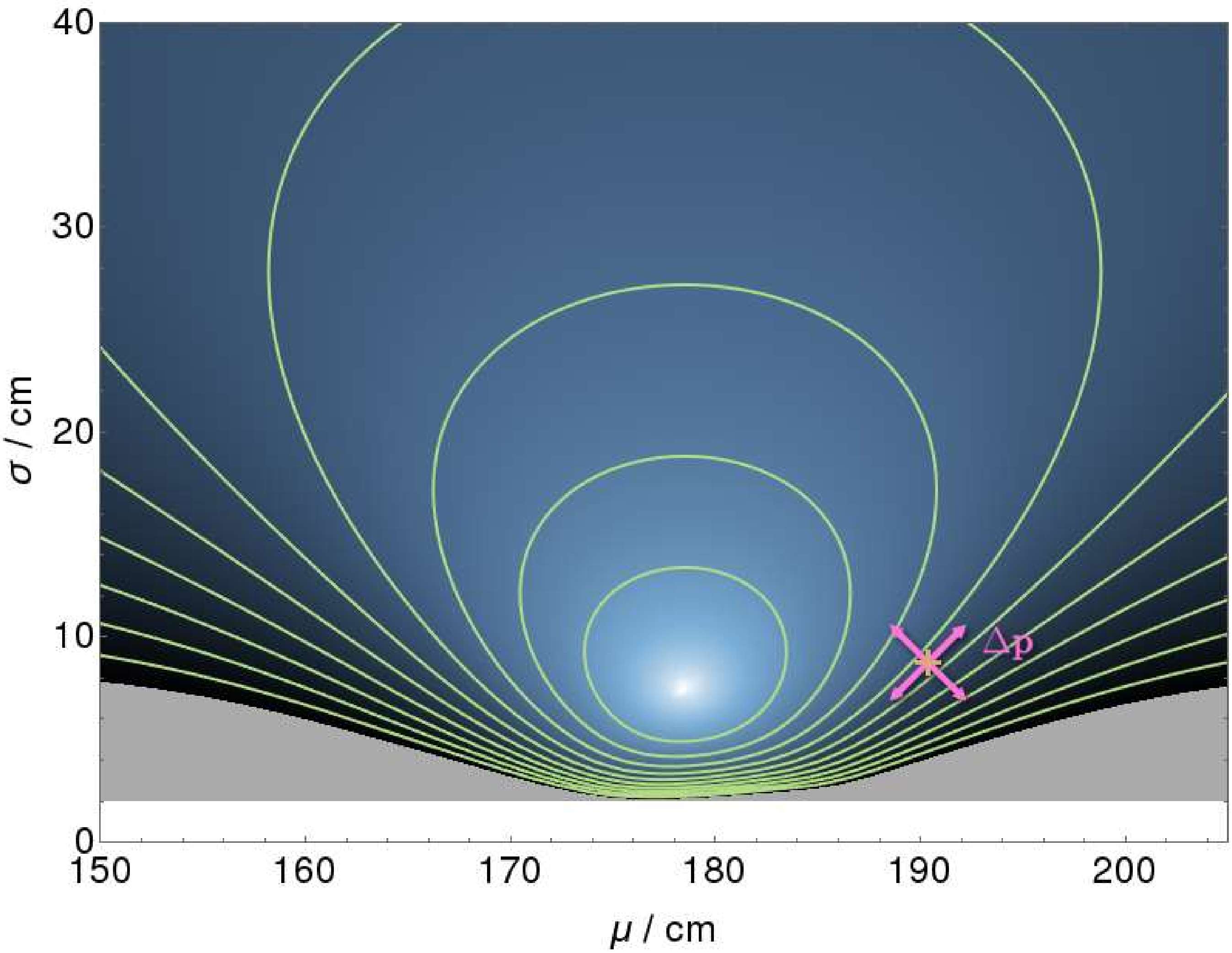
7.

We can define another vector, $\Delta\mathbf{p}$, in the same space as $\mathbf{p}$ that tells us what change can be made to $\mathbf{p}$ to get a better fit.

For example, a model with parameters $\mathbf{p}' = \mathbf{p} + \Delta\mathbf{p}$ will produce a better fit to data, if we can find a suitable $\Delta\mathbf{p}$.

The second course in this specialisation will detail how to calculate these changes in parameters, $\Delta\mathbf{p}$.

Given the following contour map,



What $\Delta\mathbf{p}$ will give the best improvement in the model?

- ( ) $\Delta\mathbf{p} = \begin{bmatrix} -2 \\ -2 \end{bmatrix}$
- (•) $\Delta\mathbf{p} = \begin{bmatrix} -2 \\ 2 \end{bmatrix}$

Correct

This direction will decrease the SSR making the fit better.

- $\Delta \mathbf{p} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$

- $\Delta \mathbf{p} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$