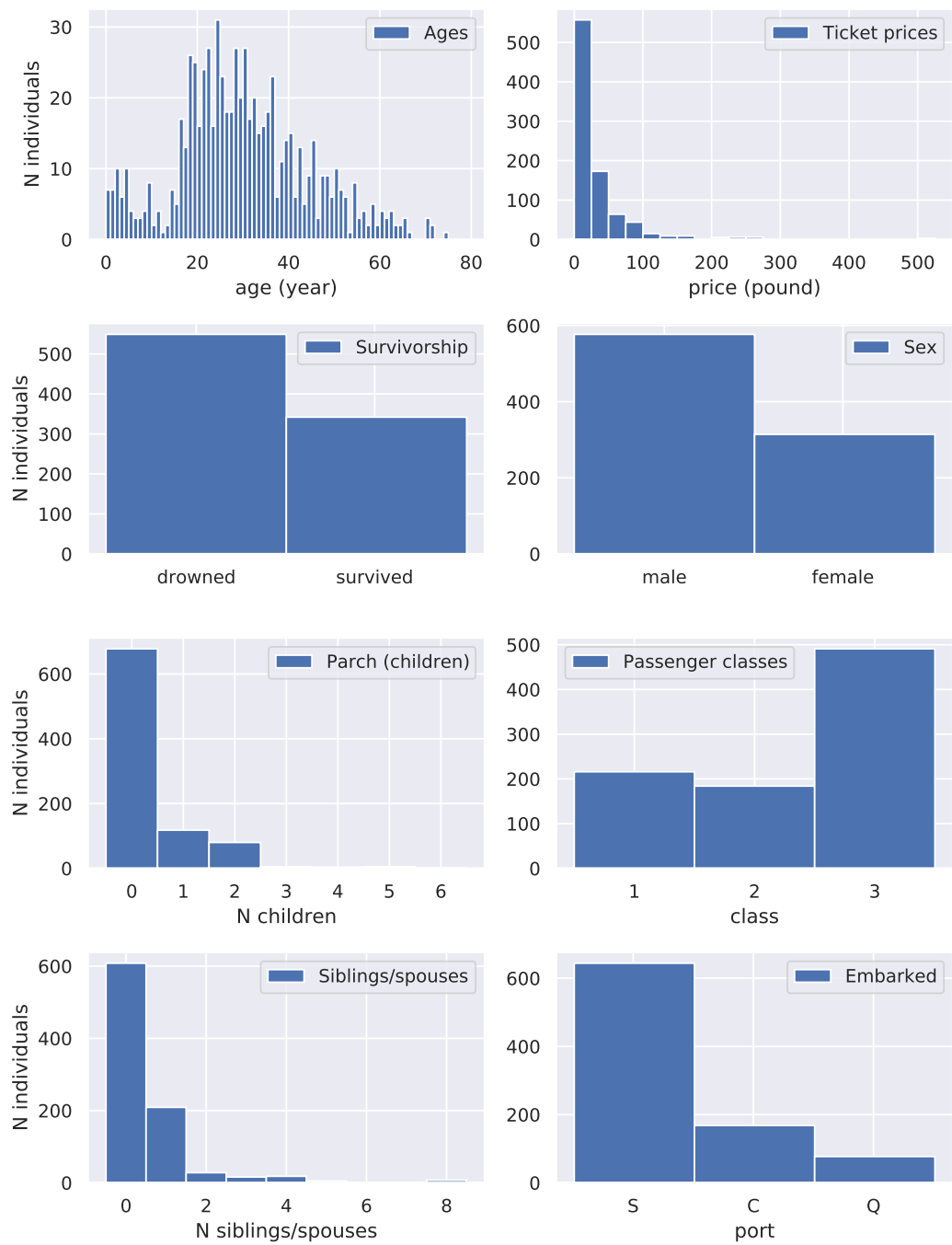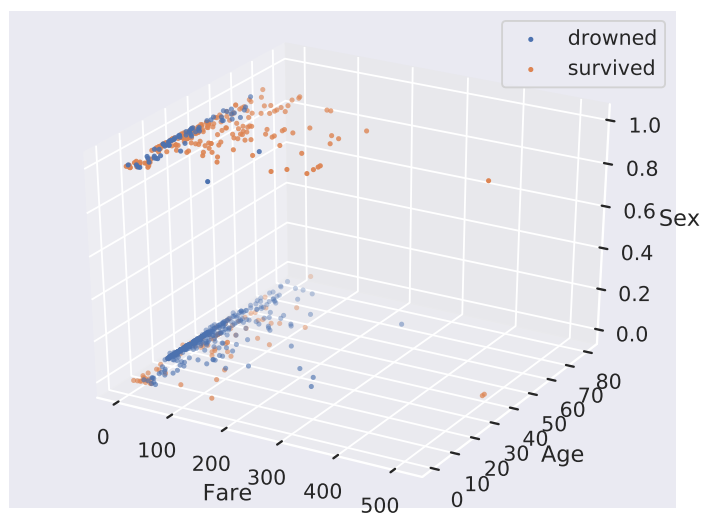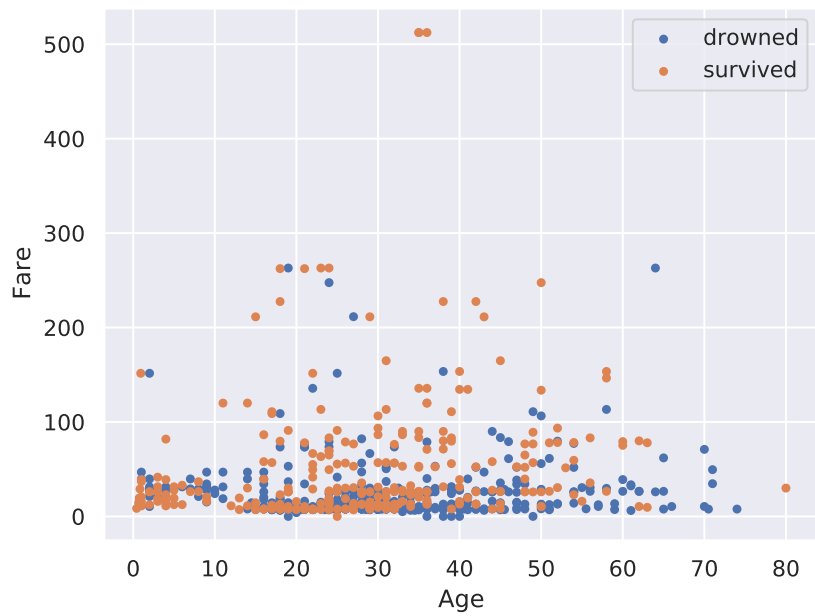# Analysis of Titanic data

Samuel Knudsen

Last update: September 10, 2021

Parameter distributions

Bayesian probability analysis

Let us look at which piece of data about an individual gives us most information about the likelihood of survival, given our dataset D.

For some choice of $x = 0, 1$, $y = 0, 1$, we have by Bayes rule:

$$\text{Prob}(\text{survived} = x | \text{sex} = y, D) \propto \text{Prob}(\text{sex} = y | \text{survived} = x, D) \cdot \text{Prob}(\text{survived} = x | D) \tag{1}$$

We can calculate

$$\text{Prob}(\text{survived} = x | D) = \frac{N_{\text{survived}=x}}{N_{\text{total passengers}}}, \tag{2}$$

and similarly

$$\text{Prob}(\text{sex} = y | \text{survived} = x, D) = \frac{N_{\text{survived}=x, \text{sex}=y}}{N_{\text{total survivors}}} \tag{3}$$

with the condition

$$\text{Prob}(\text{sex} = 0 | \text{survived} = x, D) + \text{Prob}(\text{sex} = 1 | \text{survived} = x, D) = 1 \tag{4}$$

as we only have those two options.

$$p(A, B|D) = p(A|B, D)p(B|D)$$
$$p(B, A|D) = p(B|A, D)p(A|D) \tag{5}$$

and

$$p(A, B|D) = p(B, A|D) \tag{6}$$

meaning

$$p(A|B, D) = \frac{p(B|A, D)p(A|D)}{p(B|D)} \tag{7}$$

Training a neural network to predict survival

input_layer: InputLayer	input:	[(?, 6)]
	output:	[(?, 6)]

dense: Dense	input:	(?, 6)
	output:	(?, 32)

dense_1: Dense	input:	(?, 32)
	output:	(?, 1)

Training development

loss
accuracy
val_loss
val_accuracy

epoch