Exercise 4:

1. **What is the experimental setup of choice when trying to detect overfitting?**

The experimental setup for detecting overfitting involves evaluating the model’s performance on separate datasets. This involves splitting the dataset into possibly three sets as explained in the lecture. These include:

* Training Set: This dataset is used to train the model and learn the parameters.
* Validation Set: This is used to evaluate or monitor overfitting during the training. If the losses due to validation increase and training loss decreases, there is an indication of overfitting.
* Test Set: It is used for the final evaluation of the model after it is completely trained and validated.

A comparison of the performance between training, validation and test datasets is done and if there is a large gap between training accuracy and validation/test accuracy, it implies there is an overfitting.

1. **What are methods to mitigate overfitting?**

* Increase the quality and quantity of the datasets.
* Data Augmentation to generate new training samples.
* Early stopping of the optimization/ refinement process.
* Regularization or smoothing of the model function using Lasso or Ridge Regression.

1. **What must be paid attention to when performing a train-validation split on the following datasets in the given problems?**

**(c1) Detecting pneumonia from chest x-rays. Data includes 112,120 unique images from 30,805 unique patients.**

1. Data Splitting: Ensure that all images from the same patient are in either the training or validation set, not to be able to predict for an unseen patient.
2. Stratified Sampling: use stratified sampling to maintain a proportion of pneumonia and non-pneumonia cases in training and validation data.
3. Avoid using datasets that reveal classification labels.

**(c2) Given 1000 voice recordings (single sentences) of 100 people in total from 5 German cities. The model should be able to classify the dialects of arbitrary people into one of these cities.**

1. Data Splitting: Ensure recording from the same person does not appear in both the training and validation sets.
2. Stratified Sampling: Maintain a balanced representation of recording from all cities in both sets.

**(c3) Given 1000 voice recordings (single sentences) of 100 people in total from 5 German cities. The model should be able to rate the dialects of arbitrary people from all over Germany by intelligibility.**

1. Data Splitting: Ensure diverse speakers from all cities are included in both train and validation set to capture a range of dialects.
2. Stratified Sampling: Ensure the validation dataset contains dialects from multiple cities and not one

**(c4) Given 1000 voice recordings (single sentences) of 100 people in total from 5 German cities. The model should be able to classify the person that said a given sentence.**

1. Data Splitting: Ensure no overlap of the same sentence spoken by different people in the training and validation set. Also, ensure that each speaker recording belongs entirely to the training or validation set
2. Stratified Sampling: Ensure the validation dataset contains diverse sentences to reflect real-world scenairos.

Exercise 5:

**Suppose we are estimating the regression coefficients in a linear regression model by minimizing the objective function L.**

**L(w) = RSStr(w) + λwTw**

**The term RSStr(w) = ∑(xi,yi)∈Dtr (yi − wT xi)2 refers to the residual sum of squares computed on the set Dtr that is used for parameter estimation. Assume that we can also compute an RSStest on a separate set Dtest that we don’t use during training.**

**When we vary the hyperparameter λ, starting from 0 and gradually increase it, what will happen to the following quantities? Explain your answers.**

**(a) The value of RSStr(w) will. . .**

1. remain constant. **No**
2. steadily increase. **Yes, because when the hyper parameter λ is increased, the regularization term gains more importance relative to the residual sum of squares, which forces the weight to shrink towards 0**
3. steadily decrease. **No**
4. increase initially, then eventually start decreasing in an inverted U shape. **No**
5. decrease initially, then eventually start increasing in a U shape. **No**

**(b) The value of RSStest(w) will. . .**

1. remain constant. **No**
2. steadily increase. **No**
3. steadily decrease. **No**
4. increase initially, then eventually start decreasing in an inverted U shape. **No**
5. decrease initially, then eventually start increasing in a U shape. **Yes, because the regularization term reduces overfiting by improving generations, but continuous increase result in overly simplified model and underfitting the data leading to worse performance and RSStest(w) increasing again.**