Question 1

a

b

In these regression schemes, w_0 acts as the y-intercept of the model and w_1 acts as the gradient. Therefore we would expect regression scheme (i) to have a larger gradient than (ii) since there is a smaller penalty on w_1 ($\lambda = 1$ vs $\lambda = 10$ respectively). Regression schemes (iii) and (iv) would have a smaller y-intercept than (i) and (ii) since w_0 is included in their penalties. Scheme (iv) would have a smaller gradient and y-intercept than (iii) since its penalty hyperparameter is larger ($\lambda = 10$).

From the above reasoning, regression scheme (i) matches (a) (it has the largest gradient), (ii) matches (b) (a smaller gradient than (i) but the highest w_0 since this is not in the regression's penalty term), (iii) matches (c) (a relatively small gradient y-intercept) and (iv) matches (d) ((iv) has a large penalty hyperparameter so the model would be encouraged to have a small gradient and y-intercept. (d) has the smallest y-intercept and a relatively small gradient).

c

i

The expected error when testing on the training data and $k = 1$ is 0. Assuming no contradictions in the data, then because the testing data set is the same as the training data set and $k = 1$, there is a classification region for each data point which will correctly classify the original training set.

ii

There are 10,000 data points in total. 50% of these are positive classes and 50% are negative classes. In leave-one-out cross validation on 1NN, the data point being considered will be correctly classified if its nearest neighbour (but not the data point itself, since we are leaving this out) matches its classification. Even though the distribution classifications varies depending on the region of the dataset (e.g. the left rectangle has 25% +ve and 75% -ve), each rectangle has the same number of data points so if you are choosing any data point randomly in either rectangular region, then the probability of that the data point's nearest neighbour matches its classification is 50%. Therefore the expected error is 5,000.

iii

Training a 1NN model on the left rectangle of the training data will result in 25% of the rectangle being a region that classifies a point as +ve class and 75% of the rectangle being a region that classifies a point as a -ve class. The left rectangle of the testing data is entirely -ve classes and since the data points in the test data are uniformly randomly, each data point has a 75% chance of landing in a region that classifies the point as negative and therefore correctly. The same is true for the right rectangle region. So overall, a 1NN model trained on the entire training dataset is expected to correctly classify 75% of the test data giving it an expected error of 2,500.

iv

Looking at the left rectangle of the training data and using 21-NN, you would expect the model to classify the entire region as -ve classes. This is because for each point in the left rectangle, of the 21 nearest neighbours, you would expect 25% to be +ve and 75% to be -ve so overall the region is classified as negative. When the training data is used as test data, we know that the 1,250 points that are +ve will be incorrectly classified as -ve and therefore the expected error is 1,250. This is the same for the right rectangle however the classifications are inverted. Overall, you would expect a training set error of 2,500 using a 21-NN classifier model.

d

i

The space of X is 2^p. The space of Y is 2. Therefore the size of the hypothesis class is 2^{2^p}.

ii

The probability that the version space is not $\epsilon-exhausted$ after n training examples is at most $|H|e^{-\epsilon n}$. The number of samples needed to ensure that the sample space has a 0.9 probability (P) of being $\epsilon-exhausted$ is:

$$n = \frac{1}{\epsilon}(ln|H| + ln(\frac{1}{P}))$$

Substitute in the values to yield:

$$n = \frac{1}{0.1}(ln(2^{2^{10}}) + ln(\frac{1}{0.9}))$$
$$n = 10 * (1024 * ln(2) + ln(\frac{10}{9}))$$
$$n = 7,098.88$$

Therefore at least 7,099 data points are needed to ensure the version space is $\epsilon - exhausted$ with 0.9 probability.

iii

f