Question 1

(a)

(1) RBF kernel SVM
(2) 1-nearest neighbour classifier
(3) 3-nearest neighbour classifier

(b)

$$Pr(x) = \sum_{k=1}^{k} \pi_k \, N(x|\mu_k, \Sigma_k) \quad \text{where} \quad \sum_{k=1}^{k} \pi_k = 1, 0 \le \pi_k \le 1$$

$$P(z) = \sum_{1}^{k} \pi_k$$

$$P(x|z) = \sum_{1}^{k} N(x|\mu_k, \Sigma_k)$$

(c)

if we use a 1-nearest neighbour classifier, the predicted category of the test sample is class 2, circle.

if we use a 3-nearest neighbour classifier, the predicted category of the test sample is class 1, square.

We can choose k by applying cross validation technique, choose the k which produces the minimum test error.

(d)

- **False,** because overfitting may occur.
- **False,** should use more training data instead of less.
- **False,** hyperparameter should be chosen during training progress, instead of testing.

Question 2

(a)

1. we can derive an efficient algorithm for solving the primal optimisation problem

2. we can use kernel to get optimal margin classifiers to work efficiently in very high dimensional spaces.

(b)

Because the final solution will have the requirement automatically and there is no sense in constraining the optimisation to the optimal solution. To see this, imagine some ξi is negative, then,

by setting ξi = 0, the cost is lower and none of the constraints is violated, so it is preferred. We also note due to the above reasoning we always have at least one of the ξi, ξi(gat) zero, i.e. inside the tube both are zero, outside the tube one of them is zero.

(c)

True, because the hyperplane is determined only by support vectors. Removing non-support vectors does not change the hyperplane.

False, In soft-margin case, we have slack variable which result gives more tolerance. If the dataset is not separated by a linear classifier, hard-margin and soft-margin will give different hyperplanes. Besides, the hard margin SVM might be susceptible to outliers.

Question 3

(a)

True, because decision-stump is linear classifier, the training data can NOT be separated by a linear combination of the specific combination of the specific type of weak classifiers.

(b)

False, because we only have finite weak classifiers in boosting learning. The boosting algorithm optimises each new alpha by assuming that all the previous votes remain fixed. During training, we correct the votes by selecting a certain weak classifier, which has no correlation with previous weak classifiers, it only has correlation with current dataset.

(c)

False, using weak classifier is appropriate in Adaboost. Stronger weak classifiers cannot guarantee better performance.

(d)

True, the dimensions of different features in the data may be inconsistent, and the difference between the values may be very large. Failure to process may affect the results of data analysis. Therefore, the data needs to be scaled according to a certain ratio to make it fall in a specific area

Question 4

(a)

$$L = \sum_{i=1}^{N} \| w^T x_i - \hat{y}_i \|_2^2 + \lambda \| w \|_2^2$$

$$w = (X X^T + \lambda I)^{-1} X \hat{y}^T$$

We use l2 norm regularisation to reduce overfitting.

(b)

Lasso Regression can also be used for feature selection because the coefficients of less important features are reduced to zero. This can be useful when the dataset is quite large, and you need to find which features are important.

Increasing lambda will increase bias and decrease variance.

(c)

$$\sum_i \| w^T \varphi(x_i) - y_i \|_2^2$$

$$w = \sum_i a_i \varphi(x_i) + 0$$

$$w^T \varphi(x_j) = \sum_i a_i \langle \varphi(x_i), \varphi(x_j) \rangle + \langle 0, \varphi(x_j) \rangle$$

$$= \sum_i a_i \langle \varphi(x_i), \varphi(x_j) \rangle$$

$$\sum_i \| w^T \varphi(x_i) - y_i \|_2^2$$

$$= \sum_i \left\| \sum_j a_j \langle \varphi(x_j), \varphi(x_i) \rangle - y_i \right\|_2^2$$

$$= \sum_i \| a^T K(x_i, X) - y_i \|_2^2$$

Question 5

(a)

1. Reducing computation time
2. Remove noise
3. Easy to visualise data

(b)

ISOMAP algorithm can reduce dimension of non-linear data (manifold) when the distance between two points are not meaningful if we use Euclidean distance function. In ISOMAP algorithm, Geodesic distance is used instead of Euclidean distance.

(c)

Because there is only one image captures a person wearing glass, so 'glass' object is a less important element, glass is more likely to be removed by PCA, because PCA only preserves important, common features.

(d)

False, dataset is down sampled and unsampled. For example, many outliers will be removed. If the data point is a outlier, the Euclidean distance will definitely changes.

False,

Question 6

(a)

Stochastic gradient decent works faster than normal gradient decent because the weight update computation is easier (SGD update weights only once for every sample), while standard gradient decent only update weight once after calculating all training samples.

(b)

The first layer:

$$(5 \times 5 \times 1 + 1) \times 100 = 2600$$

The second layer:

0 (max pooling has no parameters to learn)