ML Week 5 & 6

Learning objectives

- differentiate between supervised and unsupervised
- estimate performance of different supervised learning models
- implement model selection and compute evaluation measures

5.2 Forms of Supervised Learning

Supervised Learning means to train the algorithm with labelled data. The task is to estimate a function between input x and output y.

Forms of Supervised:

1. Regression problems - Estimate continuous values
2. Classification problems - Predict a category or class
3. Ranking problems - Predict relative ordering of items

Activity

Is it possible to generate classification output from regression output? How? What about generated regression value from a classification model?

We can map the continuous predictions to discrete classes to generate classification from output regression. That being said, it is more difficult to generate regression value from a classification model as we need to know the probabilities of each discrete value in each bin.

5.3 A supervised learning algorithm

Goal of supervised learning algorithm is to find a function h that can map input $X \rightarrow Y$. X and Y are two sets...

We use the **loss function** (gives error for single data point) to select a hypothesis function h from the hypothesis space H such that h is most accurate.

Empirical risk: total/average error across all training samples. So we actually choose the function of minimum risk.

$$R_{\text{emp}}(h) = \frac{1}{n}\sum_{i=1}^{n} L\big(h(x_i), y_i\big)$$

Activity

Have you ever thought about the true application of 0 – 1 loss? Write and share a simple explanation of the 0 - 1 loss function and its use?

0 - 1 loss is used for classification as 1 output presents wrong output while 0 represents good output. It is used in cases where the outputs are binary(only two possible values), for example, email is spam or not spam.

5.4 Model Complexity

Model complexity determines how well the models fits on training data and test data. A more complex model can fit very well on training data but performs poorly on test data - this is called overfitting. Underfitting is when the model performs poorly on training data as well. We need to have a model complex enough for generalisation.

5.5 Model complexity and Occam's razor

Occam's razor states that **the simplest explanation is preferable to one that is more complex**. This is a common approach in selecting a supervised learning model as it should work well in general cases.

5.6 Structural risk

We do not actually choose a model with the smallest empirical risk. In reality we have to find structural risk which is empirical risk + penalty due to complexity of model. Penalty is due to the overfitting nature of more complex models.

5.7 Classification metrics

Confusion Matrix (Contingency tables)

It represent correct and incorrect predictions in a table where the diagonal values are the correct prediction. It also shows how likely a model is going to predict one class as another. Confusion matrix is used for **classification problems.**

ROC curve

The ratio of false positive against true positive matter differently in different scenario. The ROC curve plots true positive rate vs false positive rate for algorithms.

F1 Measure

It is used when classes are imbalanced or when both false positives and false negatives matter. It uses a combination of precision and recall where precision is how many positives where true while recall is how many actual positives from all positives are correctly found.

5.8 Regression metrics

Mean square error Measures how close predictions are to the true value target

Explained Variance/R-square is the percentage of target variation

$$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}}$$
$$\mathrm{SS}_{\mathrm{res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$\mathrm{SS}_{\mathrm{tot}} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- y_i = actual value
- $\hat{y}_i$ = predicted value
- $\bar{y}$ = mean of actual values

5.9 **Partitioning data for training and testing**

Large dataset = more accurately model can learn or performance can be evaluated more accurately.

Methods of splitting data

- Sub sampling - partitions the datasets randomly into training and test sets in a specified ratio. - Might give imbalances groups
- Stratified sampling - data is split by classes then samples taken proportionally fro m each group. Class proportions preserved in test/train sets.

- Cross validation - splits the data into k smaller samples and iterates with a different sub sample selected as the test sample while the others use as training

5.10 finding best hyperparameters

hyperparameter is a parameter whose value is set before the learning process begins.
Validation is used to find the best hyperparameter from which the model will be training. Hence the model with best hyperparameters can be selected and that model will be given a final evaluation with the test data.

5.11 Effect of imbalanced classes

Imbalance classes make model training biased. This results in biased predictions (favours majority class) and misleading accuracy (high accuracy as likely to predict majority class). The solutions are either to resample the data such that the imbalances are weighted (Resampling at data level) or To adjust the costs/decision threshold at the algorithmic level.

Activity

Can you give an example of a real world problem that would probably have imbalanced data?

Rare disease diagnosis. Most people with the same conditions will not have that disease. This leads to an imbalance class that will predict with misleading high accuracy.

5.12-5.14

Multivariate regression simply means more than one feature used in the regression.

$$\hat{y} = w_0 + w_1x_1 + w_2x_2 + \cdots + w_px_p$$

We can use LinearRegression from sklearn.linear_model to plot regression line.
A **correlation matrix** is a **table of correlation coefficients** showing how strongly each pair of variables is related.
As test data size increases, regression error decreases — but it eventually **plateaus** due to noise and model limitations.

References

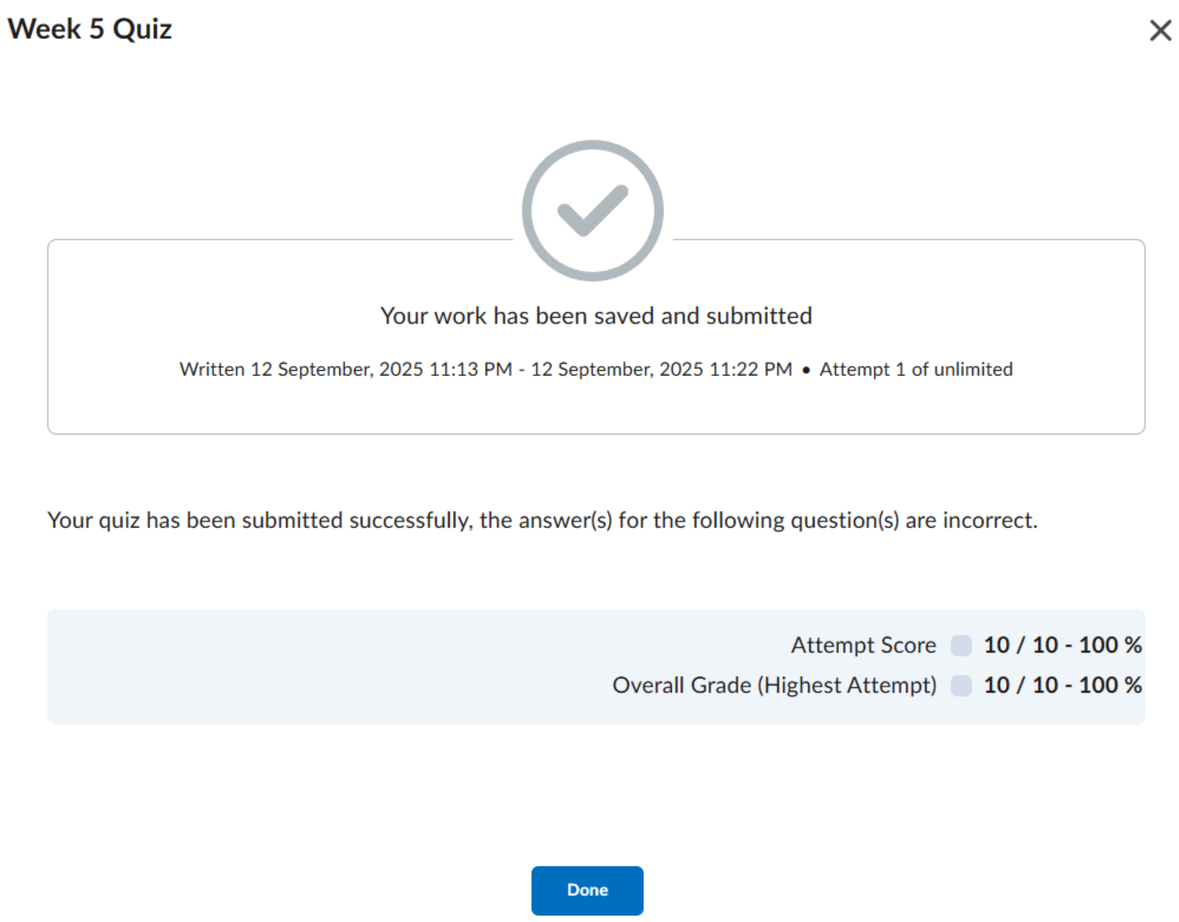
Google Developers. (n.d.) *Classification: Accuracy, Precision, Recall and related metrics*. Available at: https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall

Investopedia. (2024) *R-Squared: Definition, Calculation, and Interpretation*. Available at: https://www.investopedia.com/terms/r/r-squared.asp

Wikipedia. (2025) *Coefficient of determination*. Available at: https://en.wikipedia.org/wiki/Coefficient_of_determination

Wikipedia. (2025) *Precision and recall*. Available at: https://en.wikipedia.org/wiki/Precision_and_recall

Week 5 Quiz



Week 5 Quiz

Your work has been saved and submitted

Written 12 September, 2025 11:13 PM - 12 September, 2025 11:22 PM • Attempt 1 of unlimited

Your quiz has been submitted successfully, the answer(s) for the following question(s) are incorrect.

Attempt Score 10 / 10 - 100 %

Overall Grade (Highest Attempt) 10 / 10 - 100 %

6.2 Relevance and Covariance among features or variable

Covariance measures the amount of information a specific x_i can provide for y_i

$$\mathrm{Cov}(X,Y) \;=\; \frac{1}{n}\sum_{i=1}^{n}\bigl(x_i-\bar{x}\bigr)\bigl(y_i-\bar{y}\bigr)$$

- If **both increase together** → covariance is **positive**.
- If one increases while the other decreases → covariance is **negative**.
- If they don't show a consistent relationship → covariance is **close to zero**.

Pearson's Correlation Coefficient normalises Covariance to make it between [-1,1].

$$r \;=\; \frac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y}$$

6.3 Example of Linear Regression

Support Vector Regression tries to find a function that approximates the data within a certain margin of tolerance. It is used for non-linear relationships and when we dont want outliers to affect the regression line too much.

6.4 Linear Regression formulation

Kernel Trick give us a way to make non-linearly separable dataset linearly separable by mapping them into a higher dimensional space. The kernel trick uses a kernel function to map non-linear data into a higher dimension so that it becomes linearly separable. Hence we can use SVM on the high dimensional data.

Support Vector Machine (SVM) is used for classification. Non-linear SVM uses the kernel trick.

6.5 Linear Classification

Linear classification is when we use a linear function to separate data points into different classes - binary or multi class.

Logistic regression used for classification, not regression!

Logistic regression works by:

1. Linear predictor combines features into a single score

2. Logistic link gives a probability that an input belongs to a class

Logistic regression is preferred to least squares regression because least squares is affected by outliers.

6.6 Generalisation and complexity

Generalisation means how well a model trained on training data performs on unseen test data.

If the mean square error (MSE) on test data is low, model generalises well.

MSE on training data low but high on test data $\rightarrow$ model overfits

MSE on training data high and high on test data $\rightarrow$ model underfits

Model complexity in linear regression depends on number/type of features.

Polynomial expansion turns one existing feature into many derived features of a polynomial order d.

Polynomial expansion is useful because it makes linear regression **flexible enough** to capture non-linear patterns and interactions, while keeping it mathematically simple.

6.7 Logistic regression formulation

- Logistic regression predicts probabilities using the **sigmoid** function.
- It models the **log-odds** of belonging to a class as a linear function of the inputs.
    - **Odds** = ratio of success probability to failure probability
    - **Logit** = log of odds (can take any real value):
- Then applies a simple **threshold** (0.5) to classify.

6.8 Training a Logistic Regression model

1. Compute logistic predictor z for each x_i
2. Then for each z_i compute p_i the using the sigmoid/logistic link function
3. Use the loss function to measure difference between between predicted class and actual class.

4. Adjust w,b with gradient descent to minimise loss - optimisation. Gradient descent enables us to find lowest loss

Difference between parameters/weights and hyperparameters:

Parameters are values the model **learns from data** during training while hyperparameters are settings you choose before training.
Activity: Why do you think many believe you should run the Gradient descent with many different random initialisations?

- If you start gradient descent in a “bad spot,” it might get stuck in a poor solution.
- Running with different initialisations increases the chance of finding a **better (or global) minimum**.

6.10 Model Complexity

Overfitting happens when training model captures noise. Underfitting is the result of very simple model that doesnt capture enough information.

Bias Variance Decomposition shows that the expected loss(risk) can be broken down into 3 parts.

$$\text{Risk} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Noise}$$

Bias is the error due to the **difference between the model’s expected prediction** and the **true function** we are trying to learn.
High bias = Underfitting while low bias = overfitting

Variance measures the sensitivity of the model to training data.

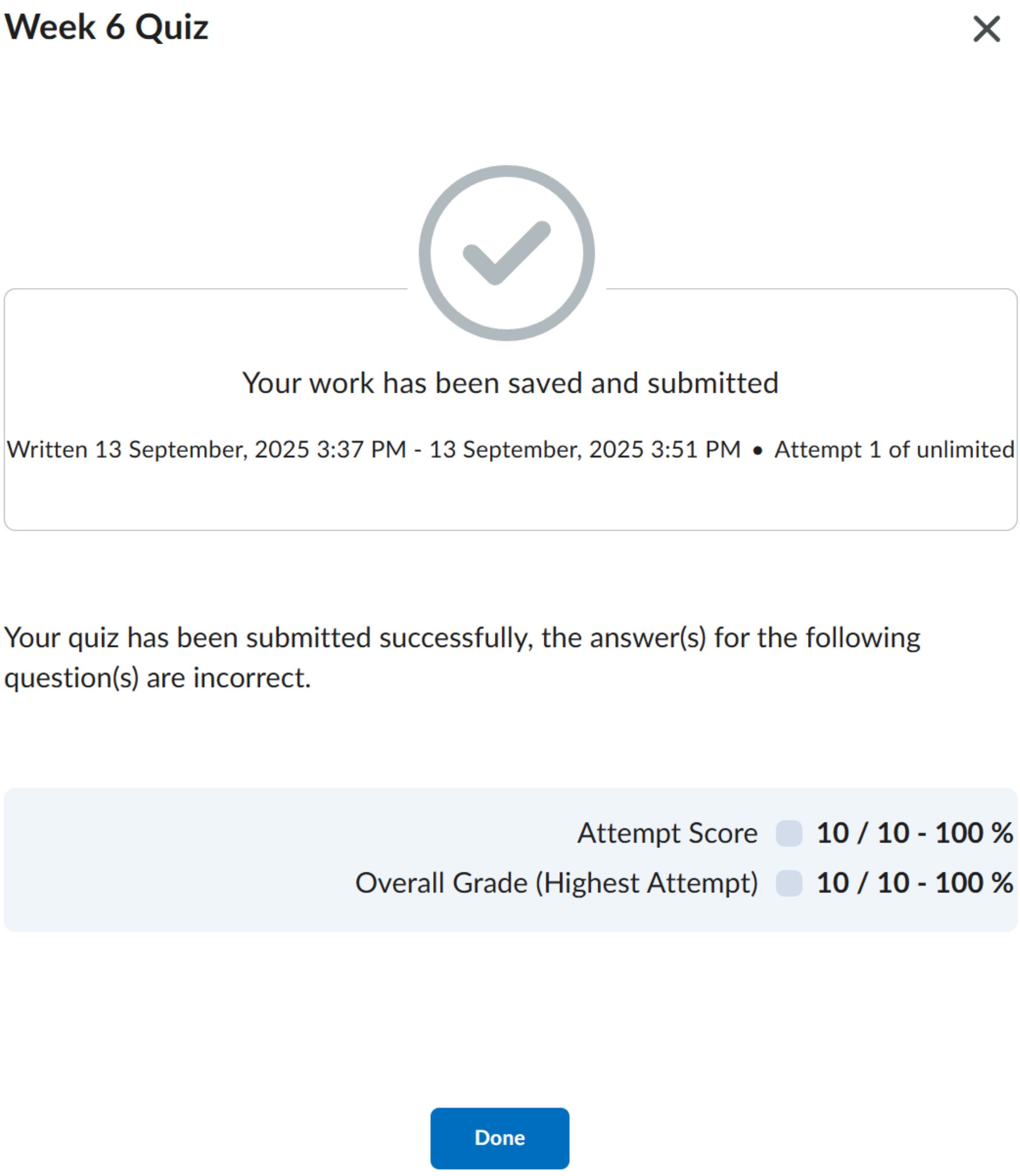
If a model has a low bias, it will naturally have a high variance as it will be sensitive to outliers.

6.11 Regularised linear models

We use a regularizer to penalty weights to reduce overfitting. regularization minimizes overfitting by giving features with more noise less weight and alpha is the parameter used to determine how much weight those features will lose.

A Regulariser prevents over-fitting by restricting the feature weights from taking very large values. L_2 regularizer is also called Ridge regularizer.

Ridge regression (L2) shrinks all coefficients smoothly but keeps every feature, while Lasso regression (L1) can shrink some coefficients exactly to zero, performing feature selection.



Week 6 Quiz

Your work has been saved and submitted

Written 13 September, 2025 3:37 PM - 13 September, 2025 3:51 PM • Attempt 1 of unlimited

Your quiz has been submitted successfully, the answer(s) for the following question(s) are incorrect.

Attempt Score	**10 / 10 - 100 %**
Overall Grade (Highest Attempt)	**10 / 10 - 100 %**

References

- Hastie, T., Tibshirani, R. & Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edn. New York:

Springer. Available at: https://hastie.su.domains/ElemStatLearn/ (Accessed: 13 September 2025).

- James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013) *An Introduction to Statistical Learning: with Applications in R*. New York: Springer. Available at: https://www.statlearning.com/ (Accessed: 13 September 2025).
- Bishop, C.M. (2006) *Pattern Recognition and Machine Learning*. New York: Springer. Available at: https://www.microsoft.com/en-us/research/people/cmbishop/prml-book/ (Accessed: 13 September 2025).
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. (2011) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825–2830. Available at: https://scikit-learn.org/stable/ (Accessed: 13 September 2025).

Reflection

I learnt a lot in week 5 and week 6 content. Unfortunately, I had to take an extension, but I am glad that I did because it allowed me time to properly learn and master the different concepts taught. I understood that logistic regression is a supervised learning technique to classify data and I was able to build one and tune its parameters for optimal perfomance. I witnessed the importance of scaling data so that the model is not biased towards larger values. I have completed my summary and the problem solving task. I must admit that my problem solving task file may be longer than expected, this is because I wanted to do every thing step by step so that to prevent confusion for myself and I elaborated my descriptions and notes as I was acquiring more knowledge while doing the task . Overall, I am satisfied and confident to tackle the next task.