Chapter 1 — The Machine Learning Landscape notes

1. How would you define Machine Learning?
ANS:

> "Machine Learning is the science (and art) of programming computers so they can learn from data."

Same section — formal definition by Arthur Samuel
Locator quote:

> "the field of study that gives computers the ability to learn without being explicitly programmed."

Same section — Tom Mitchell's formal definition
Locator quote:

> "A computer program is said to learn from experience E…"

These three appear together in the opening section of Chapter 1.

2. Can you name four types of problems where it shines?
ANS:
Machine Learning is great for:
• Problems for which existing solutions require a lot of hand-tuning or long lists of rules: one Machine Learning algorithm can often simplify code and perform bet- ter.
•Complex problems for which there is no good solution at all using a traditional approach: the best Machine Learning techniques can find a solution.
• Fluctuating environments: a Machine Learning system can adapt to new data.
• Getting insights about complex problems and large amounts of data.

3. What is a labeled training set?
ANS:

A labeled training set is:

> A dataset where each training example includes both the input data and the correct output (target).

In supervised learning:

Each example looks like this:

(x,y)(x, y)(x,y)

Where:

- x = input features (e.g., email text, house features)
- y = correct label (spam/not spam, house price)

4. What are the two most common supervised tasks?
ANS:
The most common supervised tasks are classification and regression.

5. Can you name four common unsupervised tasks?
ANS:
1. Clustering

Goal: Group similar instances together.

No labels are given.

Example from the book:

Customer segmentation

Grouping similar users

The algorithm discovers structure in data automatically.

2. Visualization

Goal: Reduce data to 2D or 3D so humans can understand it.

Used for:

Seeing structure

Identifying patterns

Spotting clusters

3. Dimensionality Reduction

Goal: Reduce the number of features while preserving important information.

Used for:

Speeding up training

Removing noise

Data compression

4. Association Rule Learning

Goal: Discover interesting relationships between attributes.

Example from the book:

If customers buy bread → they often buy butter.

This is common in:

Market basket analysis

Recommendation systems

6. What type of Machine Learning algorithm would you use to allow a robot to walk in various unknown terrains?
ANS:
Reinforcement Learning (RL)

7. What type of algorithm would you use to segment your customers into multiple groups?
ANS:
A Clustering algorithm (an Unsupervised Learning method)

8. Would you frame the problem of spam detection as a supervised learning prob- lem or an unsupervised learning problem?
ANS:
A Supervised Learning problem Specifically, a Classification problem.

9. What is an online learning system?
ANS:

An online learning system is:

A system that learns incrementally by receiving data one instance at a time (or in small batches).

Instead of training once on the whole dataset, it:

Updates the model continuously

Adapts as new data arrives

Learns progressively

How It Works:

- New data arrives
- The model updates slightly
- The old data can be discarded
- The system improves continuously

This is also called:

Incremental learning

C) When Is It Useful?

According to the book, online learning is useful when:

- You have huge datasets
- Data arrives as a stream
- The environment is changing
- You need real-time adaptation

Example:

Stock market data

Click prediction systems

Recommendation engines

D) Online vs Batch Learning

Batch Learning	Online Learning
Train once on full dataset	Train continuously
Requires retraining from scratch	Updates incrementally
Good for stable environments	Good for changing environments

E) Intuition

Batch learning = studying the whole textbook before the exam.

Online learning = learning continuously from daily practice.

10. What is out-of-core learning?
ANS:

Out-of-core learning is:

A technique for training models on datasets that are too large to fit into memory.

Instead of loading the entire dataset at once, you:

- Load a small chunk (mini-batch)
- Train the model on that chunk
- Discard it
- Load the next chunk
- Repeat

C) Why Is It Needed?

When:

- Dataset is massive (e.g., terabytes)
- RAM cannot hold the full dataset
- You must process data in parts

This is common in:

- Big data systems
- Large-scale recommendation systems
- Log analysis

D) Relationship to Online Learning

Out-of-core learning:

- Often uses online learning algorithms
- But is not necessarily real-time
- It can be done offline in batches

So:

Online learning → incremental updates
Out-of-core learning → handles data too large for memory

They are closely related but not identical.

E) Intuition

Imagine reading a 10,000-page book:
You don't memorize it all at once.
You read one chapter,
Take notes,
Move to the next.
That's out-of-core learning.

11. What type of learning algorithm relies on a similarity measure to make predictions?
ANS:

Instance-Based Learning

In instance-based learning, the system stores training examples.When a new instance arrives:

- It compares it to stored examples.
- It finds the most similar ones.
- It makes predictions based on those neighbours.
- It does not build an explicit general model.

12. What is the difference between a model parameter and a learning algorithm's hyperparameter
ANS:

Model Parameter	Hyperparameter
Learned during training	Set before training
Internal to the model	Controls the learning process
Optimized automatically	Tuned manually or via validation
Example: weights (θ)	Example: learning rate

13. What do model-based learning algorithms search for? What is the most common strategy they use to succeed? How do they make predictions?
ANS:

Model-based learning algorithms search for:

The optimal values of the model's parameters that best fit the training data.

In other words, they try to find:

θ that minimizes error

So they are searching for:

- A mathematical model
- With parameter values
- That best explains the observed data

The most common strategy is:

Minimize a cost function (or maximize a utility function).

This is typically done by:

- Defining a loss function
- Measuring how wrong predictions are
- Adjusting parameters to reduce that error

This process is called:

Optimization

Once training is complete:

1. The model has learned parameter values.

2. For a new instance, the model:

    - Applies the learned function
    - Computes the output directly

Example (Linear Regression):

After learning:

$\hat{y}=\theta 0+\theta 1$

To predict:

- Plug in new x

- Compute output

No need to compare with training examples (unlike instance-based learning).

Big Picture Flow (Model-Based Learning):

- Choose model structure
- Define cost function
- Optimize parameters
- Use learned function to predict

14. Can you name four of the main challenges in Machine Learning?
ANS:

1. Insufficient training data
2. Nonrepresentative training data
3. Poor-quality data
4. Irrelevant features

(Overfitting and underfitting are also major challenges, but more algorithm-related.)

15. If your model performs great on the training data but generalizes poorly to new instances, what is happening? Can you name three possible solutions?

ANS:

The model is overfitting.

The model has:

- Learned the underlying patterns
  AND
- Learned the noise in the training data

It memorized instead of generalizing.

Solution:

Use a less complex model:

- Reduce number of features
- Choose a simpler algorithm
- Reduce polynomial degree

Goal:

Reduce model flexibility

16. What is a test set and why would you want to use it?

ANS:

Because we want to measure:

- Generalization performance

If we evaluate only on training data:

- The model may appear very accurate.
- But it may just be memorizing (overfitting).

The test set simulates:

- New, unseen data.

17. What is the purpose of a validation set?

ANS:

A portion of the training data used to compare models and tune hyperparameters.

If we only have:

- Training set
- Test set

And we repeatedly adjust hyperparameters using the test set…

We accidentally overfit the test set.

That makes the test set unreliable.

18. What can go wrong if you tune hyperparameters using the test set?

ANS:

If you tune hyperparameters using the test set:

You overfit the test set.

Even though you are not directly training on it, you are:

- Making model decisions based on test performance
- Adjusting hyperparameters repeatedly
- Selecting the model that performs best on that specific test data

Eventually, the model becomes specialized to that test set.

19. What is repeated cross-validation and why would you prefer it to using a single validation set?

ANS:

Cross Validation:

- Split training data into **k folds**
- Train on k−1 folds
- Validate on the remaining fold
- Repeat k times (each fold used once as validation)
- Average the validation scores

Repeated Cross-Validation

This means:

Running k-fold cross-validation multiple times with different random splits and averaging all results.

So instead of:

- One single 5-fold CV

You might do:

- 5-fold CV repeated 10 times
- Then average all 50 validation scores

Because a single validation set:

- Depends heavily on how the data was split
- Can give high-variance estimates
- Might accidentally be lucky or unlucky

Repeated cross-validation:

- Uses all data for both training and validation (at different times)
- Reduces variance in performance estimates
- Produces a more stable and reliable estimate

Single Validation Set	Repeated Cross-Validation
One split only	Many different splits
High variance	Lower variance
Less reliable	More robust
Faster	More computationally expensive