Chapter 1: The Machine Learning Landscape

Study Notes & Exercise Solutions

Based on "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow"

1 Chapter Summary and Key Concepts

1.1 1. What is Machine Learning?

Machine Learning (ML) is the science of programming computers to learn from data.

- **General Definition:** It gives computers the ability to learn without being explicitly programmed.
- **Engineering Definition:** A computer program learns from *Experience E* with respect to some *Task T* and some *Performance measure P*, if its performance on T, as measured by P, improves with experience E.
- **Example:** A Spam Filter.
  - **Task (T):** Flag spam for new emails.
  - **Experience (E):** Training data (emails labeled spam or ham).
  - **Performance (P):** Accuracy (ratio of correctly classified emails).

1.2 2. Why Use Machine Learning?

- **Simplifies Code:** ML replaces long lists of hand-tuned rules with algorithms that automatically learn patterns.
- **Adapts to Change:** ML systems can adapt to new data (e.g., spammers changing "4U" to "For U") without manual intervention.
- **Solves Complex Problems:** Effective for tasks with no known algorithmic solution, such as speech recognition.
- **Data Mining:** ML can help humans learn by inspecting the patterns the algorithm discovered.

1.3 3. Types of Machine Learning Systems

1.3.1 A. Based on Supervision

1. **Supervised Learning:** Training data includes labels (solutions).
   - *Tasks:* Classification, Regression.
   - *Algorithms:* Linear Regression, SVMs, Decision Trees, Neural Networks.
2. **Unsupervised Learning:** Training data is unlabeled.
   - *Tasks:* Clustering (K-Means), Visualization (t-SNE), Dimensionality Reduction (PCA), Anomaly Detection.
3. **Semisupervised Learning:** Mix of labeled and unlabeled data.
4. **Reinforcement Learning:** An agent observes the environment, selects actions, and receives rewards or penalties to learn a *policy*.

1.3.2 B. Based on Incremental Learning

1. **Batch Learning:** Learns from all data at once (offline). Slow and resource-heavy.
2. **Online Learning:** Learns incrementally (instances or mini-batches). Good for continuous data streams or large datasets (*out-of-core learning*).

1.3.3 C. Based on Generalization

1. **Instance-based Learning:** Learns examples by heart and generalizes based on similarity.
2. **Model-based Learning:** Builds a model from examples and tunes parameters to minimize a cost function.

1.4 4. Main Challenges

- **Data Issues:** Insufficient quantity, nonrepresentative data (sampling bias), poor quality (errors/noise), irrelevant features.
- **Algorithm Issues:**
  - **Overfitting:** Model is too complex; memorizes noise. *Fix:* Regularization, more data.
  - **Underfitting:** Model is too simple. *Fix:* More complex model, better features.

1.5 5. Testing and Validating

- **Training/Test Split:** Train on one set, test on another to estimate generalization error.
- **Validation Set:** Used to compare models and tune hyperparameters (prevents overfitting to the test set).
- **Cross-Validation:** Uses multiple validation sets to avoid wasting training data.

2 Exercise Solutions

1. **How would you define Machine Learning?**
   It is the science of programming computers to learn from data to perform a task better, without explicit programming.

2. **Can you name four types of problems where it shines?**
   (1) Problems requiring complex lists of rules, (2) Fluctuating environments, (3) Complex problems like speech recognition, (4) Getting insights from large data (data mining).

3. **What is a labeled training set?**
   A training set containing the desired solution (label) for each instance.

4. **What are the two most common supervised tasks?**
   Regression (predicting a value) and Classification (predicting a class).

5. **Can you name four common unsupervised tasks?**
   Clustering, Visualization, Anomaly Detection, and Association Rule Learning.

6. **What type of Machine Learning algorithm would you use to allow a robot to walk in various unknown terrains?**
   Reinforcement Learning.

7. **What type of algorithm would you use to segment your customers into multiple groups?**
   Clustering (Unsupervised Learning).

8. **Would you frame the problem of spam detection as a supervised learning problem or an unsupervised learning problem?**
   Supervised learning (Classification).

9. **What is an online learning system?**
   A system that learns incrementally from a stream of data (sequentially).

10. **What is out-of-core learning?**
    Using online learning algorithms to train on datasets too large to fit in a computer's main memory.

11. **What type of learning algorithm relies on a similarity measure to make predictions?**
    Instance-based learning.

12. **What is the difference between a model parameter and a learning algorithm's hyperparameter?**
    A *model parameter* (e.g., slope) is internal and learned during training. A *hyperparameter* (e.g., learning rate) is external, set before training, and remains constant.

13. **What do model-based learning algorithms search for? What is the most common strategy they use to succeed? How do they make predictions?**
    They search for optimal model parameters that minimize a cost function. They make predictions by feeding new features into the model function using these parameters.

14. **Can you name four of the main challenges in Machine Learning?**
    Insufficient data, nonrepresentative data, poor-quality data, and overfitting/underfitting.

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?**
    It is **Overfitting**. Solutions: (1) Get more data, (2) Simplify the model (regularization), (3) Reduce noise in data.

16. **What is a test set and why would you want to use it?**
    Data held back from training, used to estimate the generalization error.

17. **What is the purpose of a validation set?**
    To compare models and tune hyperparameters.

18. **What can go wrong if you tune hyperparameters using the test set?**
    You overfit to the test set, leading to an optimistic error rate estimation but poor production performance.

19. **What is repeated cross-validation and why would you prefer it to using a single validation set?**
    It splits training data into complementary subsets for training/validating. It is preferred for small datasets to maximize the data available for training while still getting a good validation metric.