**Chapter 1:**

* Batch (offline) vs Online Learning
* Training on **partial set** and then using **validation set** to judge. Then train on **full training set**. Lastly you evaluate on **test set**. Also rather use **cross-validation**.
* Train on train set and then evaluate on **train-dev set**. If it shows poor performance, then it is overfitting. Otherwise now evaluate on dev set. If it shows poor performance, then it is the case of data mismatch. Lastly evaluate on test set.

**Chapter 2:**

* **RMSE** gives higher weight to large errors.
* **MAE** is for data with more outliers.
* **.corr ()** in pandas measures linear correlation only.
* Shuffling data is important as some algorithms are sensitive to the order of the data they are fed.
* One issue with **OrdinalEncoder** is that the model will assume that two nearby values are more similar than two distant values. This may work for some cases but if not, use **OneHotEncoder**.
* Why not use **.get\_dummies ()** instead of OneHotEncoder? Well, the advantage of OneHotEncoder is that it remembers which categories it was trained on.
* **Scaling data** is important as models will give more importance to data with large scales and ignore the other data.
* **Normalization** is affected by outliers while **standardization** is not.
* Before standardizing data, you may need to remove heavy tails if any. You may want to apply **square root transformation** and if the tail is heavy, then instead apply **log transformation**.
* Another approach is to **bucketize** the data according to percentiles. Then replace each feature value with the index of the bucket it belongs to.
* For **multimodal distribution**, it can also be helpful to bucketize it, but this time treating the bucket IDs as categories, rather than as numerical values. This means that the bucket indices must be encoded, for example using a OneHotEncoder. Another approach is to use **RBF**.
* **RandomizedSearchCV** is often preferable, especially when the hyperparameter search space is large over **GridSearchCV**.
* Scikit-Learn also has **HalvingRandomSearchCV** and **HalvingGridSearchCV** hyperparameter search classes. Their goal is to use the computational resources more efficiently, either to train faster or to explore a larger hyperparameter space
* **scipy.stats.randint(a, b+1):** for hyperparameters with *discrete* values that range from a to b, and all values in that range seem equally likely.
* **scipy.stats.uniform(a, b):** this is very similar, but for *continuous* hyperparameters.
* **scipy.stats.geom(1 / scale):** for discrete values, when you want to sample roughly in a given scale. E.g., with scale=1000 most samples will be in this ballpark, but ~10% of all samples will be <100 and ~10% will be >2300.
* **scipy.stats.expon(scale):** this is the continuous equivalent of geom. Just set scale to the most likely value.
* **scipy.stats.loguniform(a, b):** when you have almost no idea what the optimal hyperparameter value's scale is. If you set a=0.01 and b=100, then you're just as likely to sample a value between 0.01 and 0.1 as a value between 10 and 100

**Chapter 3:**

* **Stochastic Gradient Descent (SGD)** is good for large datasets and makes it well suited for online learning while **Support Vector Machine (SVM)** scale poorly on large datasets.
* **Accuracy** is not preferred metric for error in classification and especially for skewed datasets where one class is more prevalent than the other.
* Use **Precision-Recall curve** as a metric when the positive class is rare or when you are more interested in false positives rather than false negatives. Otherwise use **ROC curve.**
* **Multi-class Classification:**
  + Some Scikit-Learn classifiers (e.g., LogisticRegression, RandomForestClassifier, and GaussianNB) are capable of handling multiple classes natively. Others are strictly  
    binary classifiers (e.g., SGDClassifier and SVC).
  + One way is to build 1-detector, 2-detector etc. and you get the decision score from each classifier for that image, and you select the class whose classifier outputs the highest score. This is called the **one-versus-the-rest (OvR) strategy**, or sometimes **one-versus-all (OvA).**
  + Another strategy is to train a binary classifier for every pair of digits: one to distinguish 0s and 1s, another to distinguish 0s and 2s, another for 1s and 2s, and so on. You output the class that won the most duels. This is called the **one versus-one (OvO) strategy**. The main advantage of OvO is that it only needs data for the classes it is going to distinguish between.
  + For SVC, you may want to use OvO since it can’t handle large data that graciously.
* **Multi-label Classification:**
  + **KNeighborsClassifier** has the capability to do multi-label classifications.
  + If you wish to use a classifier that does not natively support multilabel classification, such as SVC, one possible strategy is to train one model per label. However, this strategy may have a hard time capturing the dependencies between the labels. For this, Scikit has **ClassifierChain** which will feed appropriate labels from the previous classifiers in chain.