Machine Learning: A Probabilistic Perspective

Everything after Chapter 4 got advanced for me LOL so this is unfinished for the most part

Chapter 1 - Introduction

- Machine learning is a set of methods that can automatically detect patterns in data and then use those uncovered patterns to predict future data or perform other kinds of decision making.
    - Best way to solve such problems is to use probability theory.
- Predictive of **supervised learning** is the goal of learning a mapping from inputs x to outputs y given a set of input and output pairs, called the training set.
- Each training input x is a D dimensional vector, where each number in that vector is called a feature or attribute.
- The outputs y can be a categorial variable from some finite set of classes. This is **classification**.
- When the outputs y are real scalar values, this is **regression**.
- The goal of **unsupervised learning** is to find interesting patterns in the given inputs.
- Binary classification is when the number of classes to classify from is 2. If C > 2, it is multi-class classification.
- The **probability distribution** over all possible labels, given the input vector x and the label y and the training set D is p(y|x, D). This represents a vector of length C, with all of the probabilities for each of the classes.
- When choosing between different models, the notation is p(y|x, D, M).
- To find out the predicted class of the model, you simply take the argmax of the probability distributions.
    - The most probable class label is called the mode of the distribution and is known as a MAP (maximum a posteriori) estimate.
- **Unsupervised learning** is where we have the task of density estimation. We want to build models of the form $p(x \mid \theta)$.
- Supervised learning is a conditional density estimation, while unsupervised learning is an unconditional density estimation. SL is conditional because we are given a training set.
- **Clustering** data is a method in unsupervised learning.
    - FIrst we determine how many clusters to create.

      $$K^* = \arg\max_K p(K|\mathcal{D}).$$

    - Then, we estimate which cluster each point belongs to. The feature of which cluster the point is a part of is a hidden or latent variable because it is not observed in the training set, but rather is something that we created. We can pick the cluster by using.

$$z_i^* = \text{argmax}_k \, p(z_i = k|\mathbf{x}_i, \mathcal{D})$$

- **Dimensionality reduction** is used to capture the essence of the data.
    - The motivation is that although data appears high dimensional, there may only be a small number of degrees of variability, corresponding to latent factors.
- Most common approach to dimensionality reduction is PCA (discuss later).
- **Matrix completion** (or imputation) is the method of inferring plausible values for missing entries in a matrix.
- **Parametric models** are ones that have a fixed number of parameters.
- The number of parameters in **nonparametric models** grows with the amount of training data.
    - K nearest neighbors is a nonparametric classifier, which looks at the K points in the training set that are closest to the input x, and counts how many members of each class are in that set.
    - Basically, the reason KNN is a nonparametric model is that the predictions depend on the size of the training set. If you think about neural networks, they have a fixed number of parameters and the predictions are solely dependent on those parameters. The parameters capture everything you know about the data.

$$p(y = c|\mathbf{x}, \mathcal{D}, K) = \frac{1}{K} \sum_{i \in N_K(\mathbf{x}, \mathcal{D})} \mathbb{I}(y_i = c) \tag{1.2}$$

where $N_K(\mathbf{x}, \mathcal{D})$ are the (indices of the) K nearest points to $\mathbf{x}$ in $\mathcal{D}$ and $\mathbb{I}(e)$ is the **indicator function** defined as follows:

$$\mathbb{I}(e) = \begin{cases} 1 & \text{if } e \text{ is true} \\ 0 & \text{if } e \text{ is false} \end{cases} \tag{1.3}$$

    - This can also be thought of an instance based learning or memory based learning.
- Curse of dimensionality comes with high dimensional data.
    - To combat this, we make assumptions about the data distribution in the form of creating a parametric model where it has a fixed number of parameters that doesn't increase with the training set.
- A Gaussian or normal distribution is one that is shaped like a bell curve.
- **Linear regression** says that the response is a linear function of the inputs.

$$y(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + \epsilon = \sum_{j=1}^{D} w_j x_j + \epsilon$$

- **Logistic regression** computes the linear combination of inputs, but also passes the output through a sigmoid function, which is necessary for the output to be interpreted as a probability.
    - If we threshold the probability, we can induce a decision rule.

$$\mu(\mathbf{x}) = \text{sigm}(\mathbf{w}^T\mathbf{x})$$

- The data is not linearly separable if there is no straight line we can draw to separate the 1s from the 0s, and thus these models will have a non zero train error.
- The lower the value for K in KNN, the more likely the model is to overfit. A lower K value signifies a complex model, while a large K underfits and is too simple.
- **Generalization error** is the error of a function that is tested on data it has never been trained on.
- A common technique to measure a model's performance is to split the training set into two pieces: A training set and a validation set which just acts as a test set.
- **Cross validation** is a technique where the training data is split into K folds, and for each fold, we train on all the folds but the k'th fold, and test on that k'th fold. The error will averaged across all of the folds.
- **Leave One Out Validation** is when you set K = N
- The **no free lunch theorem** states that there is no universally best algorithm. We use all of those previous methods (validation sets, cross validation, minimization of test error) to empirically choose the best method for our particular problem. The no free lunch theorem basically says that the performance of two models is the same if it's averaged over all possible problems. This sounds terrible, but the caveat is that we specific models are better for specific problems, and most of the time, we know the problem space we have and we're not necessarily averaging across all problems.

Chapter 2 - Probability

- Two interpretations of probability
    - Frequentist - Probabilities represent long run frequencies of events. Ex) If we flip a coin a bunch of times, it will land heads about half the time.
    - Bayesian - Probability is used to quantify uncertainty about something. Ex) 80% chance of raining tomorrow. It's basically where the probability represents how probable we think this event is.
- p(A) denotes the probability that event A is true.
- Discrete random variables are variables that can take some value from a finite set X.
- Conditional probability of A, given that B is true.

$$p(A|B) = \frac{p(A,B)}{p(B)} \text{ if } p(B) > 0$$

- Probability of an event based on a prior

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)},$$

- X and Y are unconditionally independent if p(X,Y) = p(X)p(Y). Equivalently, you can also say p(X | Y) = P(X).

    - Set of variables is mutually independent if the joint can be written as a product of marginals.
    - Basically, two events are unconditionally independent if the fact that one event happened doesn't impact the probability of another event happening.
- X and Y are conditionally independent given Z if the conditional joint can be written as a product of conditional marginals. p(X, Y | Z) = p(X|Z) p(Y|Z)
- The probability that an uncertain continuous quantity X lies between a and b can be computed as the following.

$$P(a < X \leq b) = \int_a^b f(x)dx$$

    - f(x) can be defined as the probability density function.
- Each distribution has a
    - mean (expected value)

$$\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} x \ p(x),$$

    - variance (spread of a distribution)

$$\text{var}[X] \triangleq \mathbb{E}\left[(X-\mu)^2\right] = \int (x-\mu)^2 p(x)dx$$

    - standard deviation

$$\text{std}[X] \triangleq \sqrt{\text{var}[X]}$$

- Suppose we toss a coin N times. Let X be the number of heads (anywhere from 0 to N).
    - If the probability of heads is θ, then X has a binomial distribution.

$$X \sim \text{Bin}(n, \theta)$$

$$\text{Bin}(k|n,\theta) \triangleq \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

- Suppose we toss the coin once. Let X be 0 or 1 (depending on the coin flip result), with the probability of heads as θ.
    - We say that X has a Bernoulli distribution.
    - A Bernoulli random variable is one that only has 2 outcomes.
- Encoding the states 1, 2, and 3 as (1,0,0) and (0,1,0) and (0,0,1) is a one-hot encoding.
- X can have a Poisson distribution with a parameter lambda, with the following probability mass function.

$$\text{Poi}(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$

- A couple of the most common univariate continuous distributions are
    - Gaussian (normal) distribution: Bell curve
    - Laplace distribution: Double sided exponential distribution
    - Gamma distribution: Flexible for positive real values

    - Exponential distribution: Times between events in a Poisson process
    - Beta distribution: Support over the interval [0,1]
    - Pareto distribution: Model distributions with long tails.
- Joint probability distributions are distributions on multiple related random variables.
    - It models the relationships between the variables.
- Covariance between X and Y measures the degree to which X and Y are linearly related.

$$\operatorname{cov}[X,Y] \triangleq \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]$$

- The Pearson correlation coefficient between X and Y can be defined below.

$$\operatorname{corr}[X,Y] \triangleq \frac{\operatorname{cov}[X,Y]}{\sqrt{\operatorname{var}[X]\operatorname{var}[Y]}}$$

- The multivariate Gaussian is the most widely used joint probability density function for continuous variables.
- The Kullback-Leibler divergence is a measure of the dissimilarity between two probability distributions p and q. It's also known as the relative entropy.
- Laplace's principle of insufficient reason argues in favor of using uniform distributions when there are no other reasons to favor one distribution over another.

Chapter 3 - Generative Models for Discrete Data

- Concept learning is a type of binary classification where f(x) = 1 if x is an example of the concept C and f(x) = 0 otherwise. The goal is to learn the indicator function f, which defines which elements are in the set C.
    - You can think of an example of concept learning as when a child learns to understand the meaning of a word.
- Occam's razor is the theory that the model favors the simplest hypothesis consistent with the data.
- (Couldn't understand most of this chapter to be honest)

Chapter 4 - Gaussian Models

- Multivariate Gaussian is the most widely used joint probability density function for continuous variables.
-