**CMPS242 – Machine Learning**

By: Zayd Hammoudeh

[Lecture 1 Introduction to Machine Learning 2](#_Toc495263200)

[Lecture 2 Probability Theory 3](#_Toc495263201)

[Lecture 3 Coding Theory, Relative Entropy, Convexity, and Jensen’s Inequality 5](#_Toc495263202)

# Introduction to Machine Learning

|  |  |  |  |
| --- | --- | --- | --- |
| **Univariate Polynomial Curve Fitting**   * - Learned weight vector * - Input Value * - Target Value * - Predicted value | **Overfitting** – The model is so “good” it just learns the training data.  **Techniques to Prevent Overfitting**   * Use more training data * Regularization * Select a simpler model. | **Sum of Squares/Total Error** | **Regularization**   * Penalizes large coefficient values. * Training Error always minimum when there is no regularization (i.e., ) |
| **Mean-Squared Error** |
| **Root-Mean-Square Error** |
| **Regularized Error** |

**Training Paradigms**

|  |  |  |
| --- | --- | --- |
| **Single Split Method**   * Randomly split the training set 70% training, 30% validation * Perform training on the training subset * Select the model based on their performance on the validation set. * **Advantages:**   + Simplest to implement (**easy**).   + Least computationally expensive (**cheap**). * **Disadvantage:**    + **Wastes data** as selecting best method using 30% less data.   + Highly affected by validation split profile (**unreliable**) | **-Fold Cross Validation**   * Randomly partition the training set into disjoint subsets of roughly equal size. * Perform training on training subsets and validation on the left out split. * Repeat the process holding out a different split. * Use the mean error to select the best model. * Middle approach * **Advantages:**    + Train on all the data   + Less variance than single split. * **Disadvantages**:   + **Wastes of the data** (better than single split if   + Moderate computational approach | **Leave-One-Out Cross Validation**   * Select a single record as the validation set. * Train on all other data points. * Repeat the process holding out a different single record until all have been held out. * Use the mean error to select the best model. * **Advantages**:   + Trains and validates on all data (**no data waste**) * **Disadvantages**   + Most computationally **expensive**.   + High variance (**weird behavior**) |

|  |  |  |
| --- | --- | --- |
| **CV-Based Model and Hyperparameter Selection**   * Compute Cross Validation error for all model/hyperparameter combinations. * Select the model/hyperparameter that returned the best CV error. * Train the model again using all the training data. | **Key Takeaway**: “Not enough to memorize, but need to be able to generalize.” |  |

# Probability Theory

|  |  |  |
| --- | --- | --- |
| * **Sample Space** () – Set all of possible outcomes. * **Event** – Set of outcomes, i.e., any subset of * **Probability Distribution** – Function   + **Example:** Given event , its probability distribution is: | **Axioms of Probability**   * **Axiom #1:** * **Axiom #2:** * **Axiom #3:** For an infinite sequence of disjoint events, , | **Properties of Union**   * **Disjoint Union:** * **Non-disjoint Union:** |
| **Conditional Probability**   * Amount of “A” in “B” * **Property of Intersection:** * Rewriting Conditional Probability: |

|  |  |  |
| --- | --- | --- |
| **Independence**   * Two events are independent if and only if: * Alternate Definition of Independence:   + Only holds if   + Definition implies: | **Partition**   * **Partition:** A set of **mutually exclusive** events whose union equals the sample space. * **Combining Conditional Probability and Partitioning:**    + Define the sequence as a partition   + Essentially a type of weighted average. | **Conditional Independence**   * Given a set of **independent** events: |
| **Multivariate Probability Distributions**   * **Joint Probability**: Given **at least two random variables**, it is a single composite probability distribution in the form: * **Marginal Probability**: Univariate probability distribution derived from a joint distribution: * **Conditional Probability**: Distribution conditioned on another variables value: |
| **Additional Properties of Conditional Probability**   * Each event is independent with the **sample space** as: * Independent with the **Empty Set** |
| **Bayes’ Rule**   * Relies on being part of a **partition**. * **Prior Probability:** and * **Likelihood**: and * **Posterior Probability:** |

|  |  |  |
| --- | --- | --- |
| **Rules of Probability**  **for Discrete Variables**   * **Sum Rule:** * **Product Rule:** * **Bayes Theorem:** | **Probability Densities**   * – **Probability Density Function**   + **Continuous:**   + **Discrete**: * – **Cumulative Distribution Function**   + Monotonically increasing   + **Bounded:**   + **Continuous Definition:** | **Expectations**   * **Definition:**   + **Discrete:**   + **Continuous:** * **Conditional Expectation**: * **Approximate Expectation**: Applies for both discrete and continuous random variables |

|  |  |  |
| --- | --- | --- |
| **Variances and Covariances**   * Variance: * **Covariance:** Between two random variables and * **Covariance** (Vector): | **Gaussian Distributions** | |
| **Univariate Gaussian** |  |
| **Multivariate Gaussian**   * - Dimension of and | * - **Covariance matrix** |

|  |  |  |
| --- | --- | --- |
| **Gaussian Parameter Estimation**   * Normal distribution is a likelihood function since it is given parameters (, ) * For i.i.d. samples: | **Maximum (Log) Likelihood**   * Find the parameters that maximize the likelihood * Easier if maximizing the log rather than the original function * Easier to determine with the | **MAP** – Maximizing Posterior Probability |

# Coding Theory, Relative Entropy, Convexity, and Jensen’s Inequality

|  |  |  |
| --- | --- | --- |
|  |  |  |