A Gentle Start Summary

2.1 A Formal Model

- **The Statistical Learning Framework**
- The learner input
    - Domain Set: X. eg. a set of papayas.
        - Represented by a **vector of features**.
        - Domain Points == **instances**
        - X == **instance space**
    - Label Set: Y
        - Restrict label set to be a two-element set {0, 1} or {-1, +1}.
        - eg. 1 represents papaya being tasty and 0 for not-tasty
    - Training Data: $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m))$
        - Finite sequence of X x Y
    - The Learner's Output: $h : X \rightarrow Y$
        - A prediction rule == predictor == hypothesis == classifier
        - Notation: **A(S)** ⇒ The hypothesis that a learning algorithm, A, returns upon receiving the training sequence S.
    - A simple data-generation model
        - Assume the instances are generated by some probability distribution.
        - Denote that probability distribution over X by D.
        - **!Important: Learner does not know about D**
        - Aim of Learner: figure out the labeling function $f : X \rightarrow Y$
    - Measure of Success: error
        - The probability that the predictor predicts **incorrectly**.
        - P(draw a random instance x, according to the distribution D, s.t. $h(x) \neq f(x)$)
        - $$L_{\mathcal{D},f}(h) \stackrel{\text{def}}{=} \mathop{\mathbb{P}}_{x \sim \mathcal{D}}[h(x) \neq f(x)] \stackrel{\text{def}}{=} \mathcal{D}(\{x : h(x) \neq f(x)\}).$$
            - The error of such h is the probability of randomly choosing an example x for which $h(x) \neq f(x)$.
            - The subscript (D, f) indicates that the error is measured w.r.t. the probability distribution D and the correct labeling function f.
            - AKA **generalization error, risk,** and **true error of h.**

2.2 Empirical Risk Minimization

- Goal of the learning algorithm is to MIN the error r.w.t. the **unknown** D & f.
- True error is **not** directly available to the learner.
- **Training error** is available to the learner.
    - $$L_S(h) \stackrel{\text{def}}{=} \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m},$$

- $[m] = \{1, \ldots, m\}$.
- AKA **empirical error & empirical risk.**

- Overfitting
  - Performance of the algorithm is excellent on the training set but poor on the **true world.**

2.3 Empirical Risk Minimization with Inductive Bias

- Need to guarantee that ERM has good performance w.r.t. the training set as well as over the underlying data distribution.
- What to do → **restricted search space**
  - Before seeing the data, choose a set of predictors == **hypothesis class H**
  - Each h in H is a function mapping from X to Y.
  - Choose h in H with the lowest possible error over S.
    - $$\mathrm{ERM}_{\mathcal{H}}(S) \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h),$$
    - argmin ⇒ the set of hypotheses in H that achieve the MIN value of $L_S(h)$ over H.
    - By restricting the the learner to choosing a predictor from H, we **bias** it toward a particular set of predictors ⇒ **Inductive Bias.**
    - **Not guaranteed not to overfit.**

- Finite Hypothesis Classes
  - Imposing an **upper bound** on the size of the hypotheses class ⇒ $|H| \leq$ some value.
  - **The Realizability Assumption (Definition 2.1)**
    - There exists h* in H s.t. $L_{(D, f)}(h^*) = 0$.
    - Implies that with probability 1 over random samples, S, where the instances of S are sampled according to D and are labeled by f, we have $L_S(h) = 0$.
  - **The i.i.d. assumption**
    - The examples in the training set are independently and identically distributed according to the distribution D.
    - Every x_i in S is freshly sampled according to D and then labeled according to the labeling function, f.
    - Notation: $\mathbf{S \sim D^m}$
      - m: the size of the training set S
      - $\mathbf{D^m}$: the probability over m-tuples induced by applying D to pick each element of the tuple independently of the other members of the tuple.
  - Randomness in the choice of the predictor h_S and $L_{(D, f)}(h_S)$.
    - eg. 70% of papayas are tasty. All examples in the training set is **not-tasty**. Then $\mathrm{ERM}_H(S)$ may be the constant function that labels all papayas as **not-tasty** ⇒ 70% error on the true error.
    - Denote the probability of getting a **nonrepresentative** sample by δ (delta).
    - $1 - \delta$ ⇒ **confidence parameter**.

  - Accuracy parameter

  - The quality of the prediction → (epsilon) .
  - If the generalization error is **greater than** epsilon ⇒ failure of the learner.
  - $$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}). \quad \mathcal{H}_B = \{h \in \mathcal{H} : L_{(\mathcal{D},f)}(h) > \epsilon\}.$$
  - $$M = \{S|_x : \exists h \in \mathcal{H}_B, L_S(h) = 0\} \quad \{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\} \subseteq M \ .$$
  - $$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) \ \leq \ \mathcal{D}^m(M) = \mathcal{D}^m(\cup_{h \in \mathcal{H}_B}\{S|_x : L_S(h) = 0\}).$$
- **Union Bound (Lemma 2.2)**
  - For any two sets A, B and a distribution D we have,
  - $$\mathcal{D}(A \cup B) \leq \mathcal{D}(A) + \mathcal{D}(B).$$
  - Apply the union bound to the equation above,
  - $$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) \ \leq \ \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S|_x : L_S(h) = 0\}).$$
  - $$\begin{aligned} \mathcal{D}^m(\{S|_x : L_S(h) = 0\}) &= \mathcal{D}^m(\{S|_x : \forall i, h(x_i) = f(x_i)\}) \\ &= \prod_{i=1}^{m} \mathcal{D}(\{x_i : h(x_i) = f(x_i)\}). \end{aligned}$$
  - **1 - epsilon ⇒ success rate**
  - For each individual sampling of an element of the training set we have,
  - $$\mathcal{D}(\{x_i : h(x_i) = y_i\}) = 1 - L_{(\mathcal{D},f)}(h) \leq 1 - \epsilon,$$
  - Combining the previous equation and using the inequality $\mathbf{1 - \epsilon \leq e^{-\epsilon}}$**,** we obtain that for every h in H_{Bad}
  - $$\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) \leq (1 - \epsilon)^m \leq e^{-\epsilon m}.$$
    - At most $\mathbf{(1 - \epsilon)^m}$ fraction of the training sets would be misleading.
    - **The larger m is, the smaller fraction the misleading set will be.**
  - $$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) \ \leq \ |\mathcal{H}_B| \, e^{-\epsilon m} \ \leq \ |\mathcal{H}| \, e^{-\epsilon m}.$$

- **Corollary 2.3**
  - **If** $m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$ **,**
  - If m satisfies the condition, then for any labeling function f, and for any distribution D, for which the **realizability assumption** holds, with the probability of at least 1 - δ over the choice of an i.i.d. sample S of size m, we have that for every ERM hypothesis, it holds that,

$$L_{(\mathcal{D},f)}(h_S) \leq \epsilon.$$