Gautam Ramasubramanian
Rutgers DIMACS Research

Statistical Learning Theory Section 2

Assumptions

1. **No Free Lunch Theorem:** If we cannot assume a relationship between training data and test data, then prediction is impossible

   Corollary: Measure of complexity of a hypothesis depends on the problem at hand. No conclusive measure of complexity.

2. Probabilistic Model (See Foundations #2)

3. All training examples and test examples are independant and identically distributed. (See Foundations #4)

Foundations

1. $\mathcal{X}$ is training example space, a subset of $\mathbb{R}^n$, where n is the number of features. $\mathcal{Y}$ is the label space, a subset of $\mathbb{R}$.

2. The pair $\mathcal{X} \times \mathcal{Y} = (\mathcal{X}, \mathcal{Y})$ is a random variable with *unknown* probability distribution P.

3. Goal of learning algorithm: Look at the probability distribution P and find hypothesis $g : \mathcal{X} \to \mathcal{Y}$. that can predict $\mathcal{Y}$ based on $\mathcal{X}$.

4. All the training examples and test examples in learning problem will be assumed to come from this distribution. All of the examples will come from P. This can only happen if all examples (training and test) are iid.

5. If all examples come from P, then consistant learning algorithms are possible. Consistant Learning Algorithms are such that if you feed it an unbounded number of training examples, the output hypothesis will approach optimality.

6. Unfortunately, there is no guarentee on the speed of convergence.

7. Suppose you are given $\mathcal{X}$, $\mathcal{Y}$ and P. How can we find the optimal hypothesis?

   (a) Choose a risk function - A measure of how good (or bad) a hypothesis function is at predicting?

   $$Risk = R(g) = P(g(\mathcal{X}) \neq \mathcal{Y}) = E(1_{g(\mathcal{X}) \neq \mathcal{Y}})$$

   (b) Goal is to minimize the risk. How to do that?

   (c) Standard Way - Construct Regression Function

   $$\eta(x) = E[Y|\mathcal{X} = x] = 2P(Y = 1|\mathcal{X} = x) - 1$$

   (d) Standard Way - Construct Target Function based on Regression Function

   $$t(x) = sgn(\eta(x))$$

   (e) Target function is the optimal hypothesis.

   i. Case 1 - Deterministic (Special Case) - What if the output for every x in $\mathcal{X}$ is perfectly deterministic? $P(Y = 1|\mathcal{X} = x)$ is either 1 or 0 for all x in $\mathcal{X}$. Then $\eta(x)$ will be either 1 or -1, and $t(x)$ will *never be wrong.* The risk of $t(x)$ will be 0.

   ii. Case 2 - Probabilistic (General Case) - If the output for every x in $\mathcal{X}$ is not deterministic, then there is noise. The noise level function can be defined as $s(x) = min(P(Y = 1|X = x), P(Y = -1|X = x))$. Then the risk of $t(x)$ will be $E_s(x)$.

   iii. The risk of $t(x)$ is called the Bayes Risk

8. It is clear that $t(x)$ will be the optimal hypothesis and the function is easy to get if we know P, the probability distribution for $\mathcal{X} \times \mathcal{Y}$. But the problem is that we don't know P. All we have are a finite number training examples to train our learning algorithm with. Luckily, all of them come from the distribution P, or so we assume.

9. We cannot even calculate risk for a hypothesis without P. Instead, we have to come up with our own *empirical risk function* based on the training examples given. The standard one is show below.

$$R_n(g) = \frac{1}{n}\sum_{i=1}^{n} 1_{g(\mathcal{X})\neq\mathcal{Y}}$$

10. An upside to having empirical risk function is that it converges to the regular risk function when given an unbounded number of training examples. A downside is that while n is finite, there exist functions who have a low empirical risk but high real risk. This is known as overfitting.

11. Precautions to avoid overfitting

    (a) Limit the hypotheses you are considering to a certain subset or class G. If the complexity of the hypothesis in this class is low, it will be an effective means not to overfit. However, often you will chose a class without the target hypothesis in it, so determining the hypothesis with the lowest empirical risk ($g*$) will give you a good predictor function but a suboptimal one.

    (b) Construct a sequence of classes G_d (d is the index), in order of increasing complexity. Redefine the empirical risk to include a penalty loss for a class with higher complexity (The higher the d, the higher the empirical risk). The new empirical risk will is shown below.

$$R'_n(g_n) = R_n(g_n) + penalty(d, n)$$

    (c) Set a class G that includes hypotheses with high complexities, but also have a regularizer term. A regularizer term is an expression based on parameters of the hypothesis function such that when you minimize it, you change the hypothesis g so as to reduce its complexity. Often times, it is taken to be $\|g\|^2$, if the function g is a vector in a vector space. The empirical risk includes a regularization term.

$$R'_n(g_n) = R_n(g_n) + \lambda\|g\|^2$$

    This will allow for you to minimize both the risk and complexity of the hypothesis at the same time. λ is the regularization parameter, and it intuitively is the weight or importance given to minimizing the complexity vs the risk. The right value of λ depends on the application.

    (d) Normalized Regularization

12. The main purpose of statistical learning theory is to create probabilistic bounds on the errors of hypotheses, specifically, hypotheses that a particular learning algorithm spits out.

13. 3 types of errors

    (a) Error between R_n and R.

$$R(g_n) = R_n(g_n) + B(n, G)$$

    (b) Error between g_n and $g*$

$$R(g_n) = R(g*) + B(n, G)$$

    (c) Error between g_n and t - Bayes Risk

$$R_n(g_n) = R(t) + B(n, G)$$