Machine Learning: Foundations **Fall semester, 2010**

Lecture 11: January 2, 2011

Lecturer: Yishay Mansour *Scribe: Oana Sidi, Inbal Avraham and Vera Vsevolohzky*[1]

11.1 Model Selection - Introduction

So far, each learning model determined the number of examples needed in order to learn a concept class. However, in many real cases, only a limited number of examples is available, and the learning algorithm is supposed to come up with the best hypothesis it can from the available data.

In the algorithms discussed previously, we solved accuracy problems of our hypothesis by requiring a sufficiently large number of examples, which reduces the probability of the hypothesis' error. We now deal with the case in which this cannot be done.

One example for such a case is when we have a class of an infinite $VCdim$. As we've seen in the previous lectures, if a concept class C has $VCdim = \infty$ then C is not learnable by any static learning algorithm, i.e. for any number of examples we will always be able to find a bad hypothesis which is consistent with the examples. So how can we learn such a class?

To demonstrate the problem, let's look at the concept class of a finite union of intervals on the line [0,1] (which has $VCdim = \infty$). Let us assume that we're given the following examples in the interval [0,1] :

```
+ + + - + + - - - - + - - + - - -
|                               |
0                               1
```

The target concept c_t is a set of intervals within [0,1].

Obviously, if we allow a sufficiently large number of intervals, we could easily come up with a hypothesis that is completely consistent with the data (e.g. surround every positive point with its own tiny positive interval). However, we want to predict correct classifications also for examples other than the original training set.

Adding more intervals to our hypothesis reduces the hypothesis' error on the training set, but may increase its error on new examples. For example, a positive interval surrounding

[1] Based on a scribe written by Gil Freundlich (June, 1996) and Roi Yehoshua, Ophir Gvirtzer, Zohar Ganon (May,2002).

a positive point may consist in the target concept of a 2/3 negative sub-interval and a 1/3 positive sub-interval, so adding this interval to the hypothesis can increase its "real" error. This way we may get hypotheses which are overfitted to the data, and may not generalize well to new examples.

Therefore, by Occam's Razor, in such cases we prefer simpler hypotheses which may have some error on the training set, but with high probability will predict better future observations. Returning to our example, we can make a table of the amount of errors generated by a hypothesis related to the number of intervals in the hypothesis :

Number of Intervals:	0	1	2	3	4	5	6	7	...
Number of Errors:	7	3	2	1	0	0	0	0	...

We can see that the more complex the hypothesis is, the smaller its error on the given examples. Beyond a certain complexity, all hypotheses yield 0 errors. So far, we've considered only those hypotheses which yield 0 errors on the training set, but now we're limited to the given examples and these examples may not be representative of the domain. Therefore, we want to consider simpler hypotheses, which may have some errors on the training set but generalize better to new examples.

To make the things worse, there is still the problem of noise. For a hypothesis to be completely consistent with the data, it becomes very complex. However, some of the inconsistencies in the data may be due to "noise", and the true concept may be much simpler than our consistent hypothesis. In the given example, the true concept may consist of a single interval (e.g. [0, 1/2]), and the inconsistent examples were generated due to noise. In such a case, adapting our hypothesis to the data causes the noise to get into the hypothesis.

So now we have to deal with a sample set which may be too small to accurately represent the domain, and may itself be "noisy".

In the following sections we'll consider different models for dealing with this problem. But first we'll start with building the theoretical model.

11.2 Theoretical Model

11.2.1 The Setup

Let us consider the following theoretical model.

Let H_i be the class of hypotheses, all having the same complexity-level, i (where $VCdim(H_i) = i$). Clearly, we get nested hypothesis classes :

$$H_1 \subseteq H_2 \subseteq \cdots \subseteq H_i \subseteq \cdots$$

any hypothesis of a lower complexity is included in any class of hypotheses of a higher complexity. Let $H = \cup_{i=1}^{\infty} H_i$.

For the sake of simplicity, we will assume

$$|H_i| = 2^i.$$

Let c_t be the the target concept. In contrast to our previous methods, we now do *not* assume that c_t is included within one of the H_i. The objective of the learning algorithm will be to produce a hypothesis which is "sufficiently close" to c_t (but not necessarily c_t itself).

11.2.2 Definitions

- $\epsilon(h)$ - the "real" error of h, i.e. the error of h over the entire domain X.

$$\epsilon(h) = Prob[h \neq c_t]$$

- ϵ_i - the lowest error found for any of the hypotheses in class H_i.

$$\epsilon_i = \min_{h \in H_i} \{\epsilon(h)\}$$

  Note that since $H_i \subseteq H_{i+1}$, $\epsilon_{i+1} \leq \epsilon_i$ (the probability of error decreases as the complexity level increases).

- ϵ^* - the optimal error level, i.e. the value towards which ϵ_i converges as i increases.

$$\epsilon^* = \inf_i \{\epsilon_i\}$$

  ϵ^* will not necessarily actually be obtained by any hypothesis h, but it is the lower-bound on any ϵ_i and could be approximated arbitrarily well. If for some i, $c_t \in H_i$ then $\epsilon^* = 0$.

- $\hat{\epsilon}(h)$ - the observed error, i.e. the error of hypothesis h on the given examples.

  $$\hat{\epsilon}(h) = \frac{1}{m} \sum_{x_i \in S} I(h(x_i) \neq c_t(x_i)) \ ,$$

  where S is the given set of m examples.

- $\hat{\epsilon}_i$ - the lowest observed error of any of the hypotheses in H_i.

  $$\hat{\epsilon}_i = \min_{h \in H_i} \{\hat{\epsilon}(h)\} \ .$$

11.2.3 The Problem: Overfitting

As the complexity level i of the hypothesis increases, its error on the given data $\hat{\epsilon}_i$ is reduced. Beyond complexity level m (where m is the number of examples in the given set) all the $\hat{\epsilon}_i$ will equal 0, since classes with $VCdim \geq m$ include all the possible classifications of m points, and thus one of their hypotheses must be consistent with the data.

This will happen even when the same hypothesis' real error-level, $\epsilon(h)$ (i.e. measured over the entire domain), is greater than 0, and even when $\epsilon^* >> 0$.

This happens because at high levels of complexity, the hypotheses (with the lowest levels of error on the given data) become too fitted to the given data. This phenomenon is called "overfitting".

In our case, we can not require a sufficiently large set of examples. The given data may be too small to accurately represent the entire domain. The presence of noise makes the given data even less representative of the entire domain. Thus, the overfitted hypothesis might turn out to be quite far from the true concept.

The simplistic approach for finding a good hypothesis would be to choose a hypothesis g which has the lowest value of $\hat{\epsilon}(g)$:

$$g = arg \min_{h \in \cup H_i} \{\hat{\epsilon}(h)\}$$

However, using this simplistic approach for choosing g will cause us to prefer overfitted hypotheses, because they yield the lowest $\hat{\epsilon}(h)$, namely zero observed error.

11.3 Fighting Overfitting

11.3.1 Penalty Based Models

One way to overcome the overfitting problem is to impose a complexity penalty on the complexity of the chosen hypothesis; we will then try to minimize both the observed error of the chosen hypothesis and its complexity penalty.

The chosen hypothesis g^* will, therefore, be defined as

$$g^* = arg \min_{g \in \cup H_i} \{\hat{\epsilon}(g) + Penalty(g)\} ,$$

where $Penalty(g)$ depends on the complexity of g.

We will define a measure $d(h)$ for the complexity of a hypothesis h as the lowest complexity level i such that h is found in H_i:

$$d(h) = \min_i \{h \in H_i\} .$$

Since the penalty is calculated based on $d(h)$, which is the first class in which h is found, the penalty will be the same for all hypotheses with the same complexity.

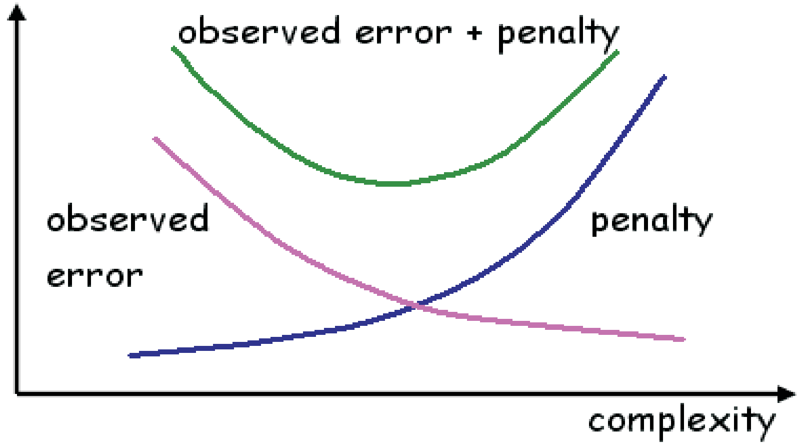

Figure 11.1: Principle of penalty based models.

Figure 11.1 shows the principle of penalty based models. As the complexity level of the hypothesis increases, its observed error is reduced but the penalty for its complexity increases. The penalty based model will try to find the minimum of the sum of the observed error and the penalty. Thus we will choose hypotheses that are not too fitted to the given examples.

11.3.2 SRM: Structural Risk Minimization

The Model

The *SRM* (*Structural Risk Minimization*) model is a penalty based model, which uses the following as the *Penalty* :

$$Penalty(h) = \sqrt{\frac{[d(h)+1]\ln(2/\delta)}{m}}\ , \tag{11.1}$$

where m is the number of examples, and δ is a confidence parameter (its meaning will be clear in the following section). This penalty defines a tradeoff between the complexity of the hypothesis and the size of the given sample. The hypothesis g^* chosen by the *SRM* model will therefore be:

$$g^* = arg \min_{g \in \cup H_i} \left\{ \hat{\epsilon}(g) + \sqrt{\frac{[d(g)+1]\ln(2/\delta)}{m}} \right\} \tag{11.2}$$

Analysis

Let h^* be the best possible hypothesis there is in $\cup H_i$, i.e., the hypothesis with the lowest actual error-level (error measured over the entire domain):

$$h^* = arg \min_{h \in \cup H_i} \{\epsilon(h)\}\ . \tag{11.3}$$

Let g^* be the hypothesis chosen by *SRM*, i.e.:

$$g^* = arg \min_{g \in \cup H_i} \left\{ \hat{\epsilon}(g) + \sqrt{\frac{[d(g)+1]\ln(2/\delta)}{m}} \right\} \tag{11.4}$$

Theorem 11.1 (*SRM* Theorem) *With probability of at least $1-\delta$ the actual error of g^* is smaller than or equal to the actual error of h^* plus twice the* SRM *complexity-penalty of h^*. Formally :*

$$\epsilon(g^*) \le \epsilon(h^*) + 2 \cdot \sqrt{\frac{[d(h^*)+1]\ln(2/\delta)}{m}} \tag{11.5}$$

Recall that by definition (of h^*) the actual error of h^* is smaller than or equal to the actual error of g^*. So, according to the *SRM* theorem, the actual error of g^* is bounded on both sides by:

$$\epsilon(h^*) \ \le\ \epsilon(g^*) \ \le\ \epsilon(h^*) + 2 \cdot \sqrt{\frac{[d(h^*)+1]\ln(2/\delta)}{m}} \tag{11.6}$$

It can be clearly seen from this inequality that the larger the number of examples (the larger m), the smaller the value of the complexity-penalty becomes, and the difference between the two hypotheses diminishes.

For the proof of the *SRM* theorem, we'll use the following claim :

Claim 11.2 *The probability that the observed error of h ($\hat{\epsilon}(h)$) will diverge from the actual error of h ($\epsilon(h)$) by more than some threshold, λ, is bounded from above:*

$$Prob\Big[|\epsilon(h) - \hat{\epsilon}(h)| \;\geq\; \lambda\Big] \;\leq\; 2e^{-\lambda^2 m} \tag{11.7}$$

Proof: This is obtained by simple application of the Chernoff Inequality. ■

Proof of *SRM* Theorem

The proof consists of two stages. First, we'll bound the error of the hypothesis in any given class H_i. Second, we'll bound the error across the classes H_i.

First stage : Bounding the error in H_i
Let g_i be the hypothesis with the lowest observed error in H_i:

$$g_i = arg \min_{h \in H_i}\{\hat{\epsilon}(h)\}$$

We want to estimate the probability of difference between the actual error and the observed error of g_i:

$$Prob\Big[|\epsilon(g_i) - \hat{\epsilon}(g_i)| \geq \lambda_i\Big]$$

(we use λ_i, because it will depend on the complexity-level i).

We cannot use Claim 11.2 directly to bound this probability, because g_i is determined according to the given sample set (and in claim 11.2 the probability is computed over all the possible sample sets).

However, we can bound this probability P by the probability that *any* hypothesis in H_i will have the difference between the actual error and observed error larger than λ_i:

$$P \leq Prob\Big[\exists h \in H_i \mid |\epsilon(h) - \hat{\epsilon}(h)| \;\geq\; \lambda_i\Big].$$

By applying the inequality of claim 11.2 we obtain:

$$Prob\Big[\exists h \in H_i \mid |\epsilon(h) - \hat{\epsilon}(h)| \geq \lambda_i\Big] \leq |H_i| \cdot 2e^{-\lambda_i^2 m}.$$

Recall that we assumed for simplicity that $|H_i| = 2^i$, so we get :

$$Prob\Big[\exists h \in H_i \mid |\epsilon(h) - \hat{\epsilon}(h)| \geq \lambda_i\Big] \leq 2^i \cdot 2e^{-\lambda_i^2 m}. \tag{11.8}$$

Let's define this upper bound (the probability of error for any hypothesis in H_i) as δ_i , i.e.:

$$\delta_i = 2^i \cdot 2e^{-\lambda_i^2 m}. \tag{11.9}$$

Solving for λ_i we get:

$$\lambda_i^2 m = \ln\left(\frac{2^{i+1}}{\delta_i}\right),$$

$$\lambda_i = \sqrt{\frac{(i+1)\ln(2) + \ln(1/\delta_i)}{m}}\ . \tag{11.10}$$

If we set the upper bound δ_i for each class H_i to $\delta_i = \frac{\delta}{2^i}$ (i.e., splitting the confidence level δ between the different classes), then we get $\delta = \sum_i \delta_i$ and thus we can use the union bound to get :

$$Prob\Big[\forall i\ \forall h \in H_i \mid |\epsilon(h) - \hat{\epsilon}(h)|\ \leq\ \lambda_i\Big] = 1 - Prob\Big[\exists i\ \exists h \in H_i \mid |\epsilon(h) - \hat{\epsilon}(h)|\ \geq\ \lambda_i\Big]$$

$$\geq 1 - \sum_i \delta_i = 1 - \delta \tag{11.11}$$

Therefore, with probability of at least $1 - \delta$,

$$|\epsilon(h) - \hat{\epsilon}(h)| \leq \lambda_i \tag{11.12}$$

for any hypothesis h in $\cup H_i$.

In this case λ_i is as follows:

$$\lambda_i = \sqrt{\frac{(i+1)\ln(2) + \ln(2^i/\delta)}{m}} = \sqrt{\frac{(2i+1)\ln(2) + \ln(1/\delta)}{m}} \tag{11.13}$$

Second stage : Bounding the error across H_i

In the previous stage, we've proved that with probability of at least $1-\delta$, for any hypothesis h in $\cup H_i$,

$$|\epsilon(h) - \hat{\epsilon}(h)| \leq \lambda_i$$

where λ_i depends on the complexity level i of h.

Among other hypotheses, this is also true for h^* and g^*, which leads to the following:

$$\hat{\epsilon}(h^*) \leq \epsilon(h^*) + \lambda_i \tag{11.14}$$

$$\epsilon(g^*) - \lambda_j \leq \hat{\epsilon}(g^*) \tag{11.15}$$

where,

- $i = d(h^*)$, i.e. i is the complexity level of h^*.
- $j = d(g^*)$, i.e. j is the complexity level of g^*.

Let's define P_i, P_j as the *SRM* complexity-penalties for h^* and g^*, respectively. Therefore, from the definition of the *SRM* model we get :

$$\hat{\epsilon}(g^*) + P_j \leq \hat{\epsilon}(h^*) + P_i \tag{11.16}$$

(otherwise the *SRM* model would not have chosen g^*).

From the three inequalities (11.14), (11.15) and (11.16) we get:

$$\epsilon(g^*) - \lambda_j + P_j \leq \epsilon(h^*) + \lambda_i + P_i \tag{11.17}$$

and therefore,

$$\epsilon(g^*) \leq \epsilon(h^*) + \lambda_i + P_i + \lambda_j - P_j\ . \tag{11.18}$$

Now, from the definition of the penalty-value for complexity-level j we get :

$$P_j = \sqrt{\frac{[j+1]\ln(2/\delta)}{m}} \geq \sqrt{\frac{(2j+1)\ln(2) + \ln(1/\delta)}{m}} = \lambda_j \tag{11.19}$$

(for $\delta \leq 0.5$). Hence the penalty is greater than the actual divergence with probability of at least $1 - \delta$.

Now we can return to inequality (11.18) and get :

$$\epsilon(g^*) \leq \epsilon(h^*) + \lambda_i + P_i + (\lambda_j - P_j) \leq \epsilon(h^*) + \lambda_i + P_i \tag{11.20}$$

(because $\lambda_j - P_j \leq 0$).

Since for P_i, too, the relation $\lambda_i \leq P_i$ holds, we get :

$$\epsilon(g^*) \leq \epsilon(h^*) + \lambda_i + P_i \leq \epsilon(h^*) + P_i + P_i = \epsilon(h^*) + 2 \cdot P_i \tag{11.21}$$

which proves the *SRM* theorem. □

In Practice

In theory, we can consider all classes H_i in $\cup H_i$, even when i goes to infinity, and search for the best hypothesis g^*. However, in practice, we have to stop at some level of complexity, i_{max}; what should this level be? By the time i reaches m, the hypotheses are complex enough to reach an observed error of zero, $\hat{\epsilon}_m = 0$ (we assumed $VCdim(H_i) = i$). Beyond this level there is no need to search, because the observed error will remain 0, and only the complexity penalty will rise (therefore we will never choose those hypotheses, because we have ones with lower complexity penalties).

11.4 Cross Validation

The Model

The *SRM* model tackled the overfitting problem by imposing a complexity penalty on the "price" of a hypothesis, which will steer us to prefer simpler hypotheses rather than complex ones. The model shows that the chosen hypotheses will not be too fitted to the given examples; this will enable the hypotheses to correctly classify also new examples which were not used in the learning process.

The *Cross Validation* method does not change the price of the hypothesis, but leaves it to be the observed error, $\hat{\epsilon}(h)$. To overcome the overfitting problem, the *Cross Validation* splits the given set of examples, S, into two sets, S_1 and S_2. The set S_1, is used as the training sample set in the learning process; this yields for each H_i some hypotheses which are estimated to be the best according to the training set. The examples of the other set, S_2, are then used as a test set, to test the error of the chosen hypotheses on the "new" examples. The chosen hypothesis will be the one with the lowest observed error on the "test" set, S_2. Therefore, *Cross Validation* deals with the overfitting problem by estimating how bad a hypothesis is when learning new examples (how tightly fit it is to the training sample set).

We denote by γ the fraction ($0 < \gamma < 1$) of the original set, S, which is reserved as the test set, S_2. Therefore, if the original set S contains m examples, then the test set S_2 contains γm examples, and the training set S_1 contains $(1-\gamma)m$ examples. Usually γ will be small, because after choosing the best hypotheses according to the training set S_1, the number of candidate hypotheses is reduced and thus we need less examples to choose the best one of the selected hypotheses according to the test set S_2.

We will divide the algorithm into two stages:

1. **Learning from S_1**:
   From each hypotheses class, H_i, we choose the best hypothesis g_i according to the training sample set S_1, i.e. the hypothesis which has the lowest observed error on S_i.

   $$g_i = arg \min_{h \in H_i} \{\hat{\epsilon}_1(h)\}$$

   where $\hat{\epsilon}_1(h)$ is the observed error of hypothesis h on the training sample set S_1.

   This will yield a set of hypotheses, G, with one hypothesis for each class H_i.

   Note that, for practicality's sake, we take $1 \leq i \leq m$; the best hypotheses from classes with complexity greater than or equal to m will have already become completely fitted to the data, and yield $\hat{\epsilon}_1(h) = 0$. We will therefore assume that $|G| = m$. (Alternatively, we will include g_i in G only if $\hat{\epsilon}_1(g_i) < \hat{\epsilon}_1(g_{i-1})$, and this can happen at most m times.)

2. **Testing on** S_2:
From G we now choose the hypothesis which has the lowest error on the test sample set S_2.

$$g^* = arg \min_{g_i \in G} \{\hat{\epsilon}_2(g_i)\}$$

where $\hat{\epsilon}_2(h)$ is the observed error of hypothesis h on the test sample set S_2.

Analysis

The analysis will show that *Cross Validation* approximates *SRM*.

Theorem 11.3 *Let* $\epsilon_{CV}(m)$ *be the error of* Cross Validation *(CV) on* m *samples, and* $\epsilon_A(m)$ *be the error of some algorithm* A *on* m *samples.*

With probability $1 - \delta$*,*

$$\epsilon_{CV}(m) \leq \epsilon_A((1-\gamma)m) + 2 \cdot \sqrt{\frac{\ln(2m/\delta)}{\gamma m}}$$

Proof:

We'll assume that algorithm A chooses the best hypothesis g_k from some class H_k, by learning from the training sample set S_1. This is the case for the *SRM* algorithm, since as we've seen the complexity penalty is the same for all hypotheses in H_k, thus if the algorithm chooses an hypothesis g_k from class H_k we're guaranteed that g_k is the best hypothesis in H_k (the hypothesis with the lowest observed error on S_1). The penalty only chooses between the best hypotheses in different classes.

Cross Validation, however, may choose a different hypothesis, g_j, because it better classifies the examples from the test set S_2. Note that g_j belongs to H_j, which is a different class from H_k. As shown before, g_k was the best in class H_k, and therefore inserted into G; it is the only member of H_k in G. If g_j is not g_k, then it comes from a different class.

First, we would like to bound the difference between the observed error and the actual error (both on the test set S_2) of any hypothesis g_i in G (the set of best hypothesis from each H_i, as chosen by *Cross Validation*).

We will use claim 11.2 from the analysis of *SRM*, and state that the probability that a hypothesis g_i in G will have the difference between its observed error on S_2 and its actual error larger than λ, is bounded as follows:

$$Prob\Big[|\epsilon(g_i) - \hat{\epsilon}_2(g_i)| \geq \lambda\Big] \leq 2 \cdot e^{-\lambda^2 \gamma m} \tag{11.22}$$

where $\hat{\epsilon}_2(g_i)$ is the observed error on the test set S_2 for hypothesis g_i.

Therefore, we can bound the probability that *any* hypothesis in G will have the difference between its actual error and observed error on S_2 larger than λ by:

$$Prob\Big[\exists g \in G \mid |\epsilon(g) - \hat{\epsilon}_2(g)| \geq \lambda\Big] \leq |G| \cdot 2e^{-\lambda^2\gamma m} \tag{11.23}$$

Since $|G| = m$ we get:

$$Prob\Big[\exists g \in G \mid |\epsilon(g) - \hat{\epsilon}_2(g)| \geq \lambda\Big] \leq m \cdot 2e^{-\lambda^2\gamma m} \tag{11.24}$$

If we set this upper bound to δ, we get:

$$\delta = m \cdot 2e^{-\lambda^2\gamma m}$$

Solving for λ leads to:

$$\lambda^2\gamma m = \ln(2m/\delta)$$

$$\lambda = \sqrt{\frac{\ln(2m/\delta)}{\gamma m}} \tag{11.25}$$

Thus we get :

$$Prob\Big[\exists g \in G \mid |\epsilon(g) - \hat{\epsilon}_2(g)| \geq \lambda\Big] \leq \delta \tag{11.26}$$

Therefore, with probability $1 - \delta$ we have for any g_i in G :

$$|\epsilon(g_i) - \hat{\epsilon}_2(g_i)| \leq \lambda. \tag{11.27}$$

From this we get:

$$\epsilon(g_j) - \lambda \leq \hat{\epsilon}_2(g_j) \tag{11.28}$$

and

$$\hat{\epsilon}_2(g_k) \leq \epsilon(g_k) + \lambda. \tag{11.29}$$

Since *Cross Validation* preferred g_j to g_k, we know:

$$\hat{\epsilon}_2(g_j) \leq \hat{\epsilon}_2(g_k). \tag{11.30}$$

Now, from the three inequalities 11.28, 11.29 and 11.30 we can get:

$$\epsilon(g_j) - \lambda \leq \hat{\epsilon}_2(g_j) \leq \hat{\epsilon}_2(g_k) \leq \epsilon(g_k) + \lambda \tag{11.31}$$

$$\epsilon(g_j) \leq \epsilon(g_k) + 2\lambda. \tag{11.32}$$

We note that:

- $\epsilon(g_j)$ is $\epsilon_{CV}(m)$, the error of *Cross Validation* when learning from m examples.

- $\epsilon(g_k)$ is $\epsilon_A((1-\gamma)m)$, the error of algorithm A when learning from $(1-\gamma)m$ examples.
- $\lambda = \sqrt{\frac{\ln(2m/\delta)}{\gamma m}}$

Thus we get:

$$\epsilon_{CV}(m) \ \leq \ \epsilon_A((1-\gamma)m) + 2 \cdot \sqrt{\frac{\ln(2m/\delta)}{\gamma m}} \ , \tag{11.33}$$

which proves theorem 11.3. □

In Practice

The analysis showed that *Cross Validation* chooses a hypothesis which, with probability $1-\delta$, is within a very close range (2λ) to the hypothesis chosen by *SRM.*

Also, *Cross Validation* tests with S_2 only a small subset of the hypotheses ($|G| = m$), only the ones that are best (in their classes) in relation to S_1.

However, the analysis compared how *Cross Validation* learns from m examples with how algorithm A learns from $(1-\gamma)m$ examples. This note is important, because there are many cases in which reducing the number of examples might significantly increase the learning error. In such cases, $\epsilon_A((1-\gamma)m)$, A's learning error from $(1-\gamma)m$ examples, is only a weak boundary on $\epsilon_{CV}(m)$, the error of *Cross Validation*'s learning from m examples, and therefore is not very useful. This is typical in cases where there is a "turning-point", a number of examples beyond which learning becomes easy (few errors), and under which learning is difficult (many errors).

It is advisable, therefore, to select a γ which will not cause such a difference in the error-level of A (estimate A's learning error as a function of m, and select a good γ).

11.5 MDL: Minimum Description Length

Consider the typical formulation of a definition in everyday speech. We prefer to define new concepts by comparing them to a simple similar concept, and describing the differences between the two. For example, we would define a car as a wagon with a motor; this is much simpler (shorter) than defining it as a four wheeled land vehicle propelled by the force of an engine. This pattern of concept definition is useful because a great deal of the information is conveyed concisely in the simple concept (here "wagon"), and the list of corrections is manageable (here "with a motor"). In many contexts, the length of the description is a criteria for choosing a description; for example, we may want to transmit the definition, and strive to minimize the length of transmission.

Of course, there is a balance to maintain between the complexity of the concept and the amount of corrections. The more complex the concept described by the hypothesis, the longer the description will be, but the description of the exceptions will be shorter. Conversely, the simpler the concept, the shorter its own description, but the longer the description of the exceptions.

The *Minimal Description Length* model proposes that when learning a new concept, we do not have to aim for the most accurate hypothesis; we can represent an accurate hypothesis by describing a simple hypothesis, which is similar to the concept but is not accurate, and supplying a list of corrections (examples which should be classified differently than the described hypothesis does). The overall description-length is the sum of the description length of the hypothesis and the corrections.

MDL proposes to find a hypothesis which minimizes the overall description length:

$$g^* = arg \min_{h \in \cup H_i} \left\{ size(corrections(h)) + size(h) \right\} ,$$

where $size(h)$ is the description size of hypothesis h, and $size(corrections)$ is the description size of the examples misclassified by h.

This model can be viewed as a penalty-based model : $size(corrections(h))$ is a function of the observed error of h, and $size(h)$ acts as the complexity-penalty. The chosen hypothesis g^* can be described as:

$$g^* = arg \min_{h \in \cup H_i} \left\{ f(\hat{\epsilon}(h)) + Penalty(h) \right\} ,$$

where $f(\hat{\epsilon}(h))$ corresponds to the description size of the observed error of h, and $Penalty(h)$ is the description size of h.

This model is related to the *MAP - Maximum A Posteriori* approach. In *MAP* we choose the hypothesis with the maximum a-posteriori probability given the data. The a-posteriori probability of a hypothesis h given the data D can be computed by Bayes rule as follows :

$$Pr[h|D] = \frac{Pr[D|h] \cdot Pr[h]}{Pr[D]}$$

The chosen hypothesis according to the *MAP* approach is :

$$g^* = arg \max_{h \in \cup H_i} Pr[h|D]$$

Since $Pr[D]$ doesn't depend on h, we can get :

$$g^* = arg \max_{h \in \cup H_i} Pr[D|h] \cdot Pr[h]$$

By taking log, we get :

$$g^* = arg \max_{h \in \cup H_i} log(Pr[D|h]) + log(Pr[h])$$

And by changing the sign :

$$g^* = arg \min_{h \in \cup H_i} log(\frac{1}{Pr[D|h]}) + log(\frac{1}{Pr[h]})$$

The expression $\frac{1}{Pr[D|h]}$ corresponds to the description size of the corrections of h in relation to the data D and the expression $\frac{1}{Pr[h]}$ corresponds to the description size of h. Hence, the problem of finding a minimum description length hypothesis is equivalent to the problem of finding a maximum a-posteriori hypothesis.

11.6 Regularization Framework

We've seen before a special case of regularization, in the SVM methodology ($\min \frac{1}{2}\|w\|^2$).

Linear Regression

A linear regressor is a mapping $x \mapsto w \cdot x$, where we assume that the instance space is a vector space (i.e. x is a vector) and the prediction is a linear combination of the instance vector x. The problem of learning a regression function with respect to a hypothesis class of linear predictors is called *linear regression.*

Formally, let $(x_1, y_1), (x_2, y_2)...(x_m, y_m)$ be a sequence of m training samples, where for each i we have $x_i \in \mathbf{R}^n$ and $y_i \in \mathbf{R}$ (notice y_i takes continuous values, not a classification problem). Consider class of linear predictors

$$\mathcal{H} = \{x \mapsto w \cdot x : w \in \mathbf{R}^n\}$$

We shall find the best hypothesis with respect to the squared loss

$$\sum_{i=1}^{m} \frac{1}{2}(w \cdot x_i - y_i)^2$$

To solve the above problem we calculate the gradient of the objective function and compare it to zero. That is, we need to solve

$$\sum_{i=1}^{m}(w \cdot x_i - y_i) \cdot x_i = 0 \text{ (notice that } x_i \text{ is a } n \text{ dimensional vector)}$$

We can rewrite the above as the problem $Aw = b$ where

$$A = (\sum_{i=1}^{m} x_i x_i^T) \text{ and } b = \sum_{i=1}^{m} y_i x_i$$

If the training instances span the entire space $\mathbf{R}^n$ then A is invertible and the solution is

$$w = A^{-1} b$$

If the training instances do not span the entire space then A is not invertible. Nevertheless, we can always find a solution to the system $Aw = b$ because b is in the range of A (the proof is omitted here).

Problem

The least-squares solution we presented before might be highly non-stable - namely, a slight perturbation of the input causes a dramatic change of the output - leading again to overfitting.

Consider for example the case where $\mathcal{X} = \mathbf{R}^2$ and the training set contains two examples where the instances are $x_1 = (1, 0)$ and $x_2 = (1, \epsilon)$ and the targets are $y_1 = y_2 = 1$. The problem becomes

$$P = \min_w \left[\frac{1}{2}(w_1 - 1)^2 + \frac{1}{2}(w_1 + \epsilon w_2 - 1)^2 \right]$$

By vanishing the gradients we get the following system of equations:

$$\begin{cases} \frac{\partial P}{\partial w_1} = w_1 - 1 + w_1 + \epsilon w_2 - 1 = 0 \\ \frac{\partial P}{\partial w_2} = \epsilon(w_1 + \epsilon w_2 - 1) = 0 \end{cases}$$

With the solution:

$$\begin{cases} w_1 = 1 \\ w_2 = 0 \end{cases}$$

Now, lets repeat the above calculation with the slight change in target: $y_1 = 1 + \epsilon$.

$$P = \min_w \left[\frac{1}{2}(w_1 - (1 + \epsilon))^2 + \frac{1}{2}(w_1 + \epsilon w_2 - 1)^2 \right]$$

$$\begin{cases} \frac{\partial P}{\partial w_1} = w_1 - (1 + \epsilon) + w_1 + \epsilon w_2 - 1 = 0 \\ \frac{\partial P}{\partial w_2} = \epsilon(w_1 + \epsilon w_2 - 1) = 0 \end{cases}$$

The solution is:

$$\begin{cases} w_1 = 1 + \epsilon \\ w_2 = -1 \end{cases}$$

That is, for the same instances, a tiny change in the value of the targets makes a huge change (from $w_2 = 0$ to $w_2 = -1$) in the least squares estimator.

A problem suffering from such instability is also called an ill-posed problem. A common solution is to add regularization. We shall see two kinds of regularization.

1. **Ridge Regression**

   We regularize by adding $\|w\|^2$ to the optimization problem, namely, to define the estimator as

   $$argmin_{w \in \mathbf{R}^n} \left(\frac{\lambda}{2} \|w\|_2^2 \right) + \hat{SQ}_w \tag{11.34}$$

   where λ is the regularization parameter and $\hat{SQ}_w = \sum_{i=1}^{m} \frac{1}{2}(w \cdot x_i - y_i)^2$ is the square loss.

   To solve Eq.(11.34) we again compare the gradient to zero and obtain the set of linear equations:

   $$(\lambda I + A)w = b$$

   Since A is positive semi-definite, the matrix $(\lambda I + A)$ has all its eigenvalues bounded below by λ. Thus, all the eigenvalues of $(\lambda I + A)^-1$ are bounded above by $1/\lambda$ which guarantees a stable solution.

2. **Lasso**

   Another form of regularization is the l_1 norm. The resulting estimator is called Lasso:

   $$argmin_{w \in \mathbf{R}^n} (\lambda \|w\|_1) + \hat{SQ}_w \tag{11.35}$$

   where again λ is the regularization parameter and $\hat{SQ}_w = \sum_{i=1}^{m} \frac{1}{2}(w \cdot x_i - y_i)^2$ is the square loss.

   While there is no closed form solution for the Lasso problem, it can still be solved efficiently by an of-the-shelf convex optimization method. In particular, we can apply the stochastic sub-gradient method for the Lasso problem.

11.6.1 Generalization error bound for SQ class

Consider the following optimization problem:

$$\min \sum \Psi_i^2$$

$$\text{s.t. } \Psi_i = w \cdot x_i - y_i \text{ and } \|w\|_2^2 \leq \Lambda^2.$$

What is the solution's generalization ability? Let's assume that

$$\|x\| \leq R \text{ and given } \|w\| \leq \Lambda$$

We are interested in bounding the Rademacher complexities of the SQ class so we can obtain generalization error bound. To this end we first state the following Lemma (proof ommited):

Lemma 11.4 *[Ledoux-Talagrand] Let $\Phi_i : \mathbf{R} \to \mathbf{R}$ be L-Lipschitz functions with parameter L $\forall i = 1, ..., n$,i.e. $|\Phi_i(a) - \Phi_i(b)| \leq L|a - b|, \forall a, b \in \mathbf{R}$. Then*

$$\hat{R}_S(\Phi \circ H) = \mathbf{E}_\sigma \left[\sup_{h \in H} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i \Phi_i(h(x_i)) \right| \right] \leq L\mathbf{E}_\sigma \left[\sup_{h \in H} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(x_i) \right| \right] = L\hat{R}_S(H)$$

Back to our optimization problem, let's assume that $\Psi_i = w \cdot x_i - y_i$ are bounded from above by M. Denote by $\Phi(\Psi) = \Psi^2$ then the function Φ is 2M-Lipschitz and by applying Lemma 11.4 we get:

$$\hat{R}_S(SQ) \leq 2M\hat{R}_S(Linear) \leq 2M\frac{R\Lambda}{\sqrt{m}} \text{ (last inequality proven in home assignment)}$$

Then the following bound holds with high probability $\geq 1 - \delta$.

$$\mathbf{E}[SQ] \leq \hat{SQ}(S) + 2\hat{R}_S(SQ) + O\left(\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right)$$

In conclusion, by bounding $\|w\| \leq \Lambda$ we guarantee generalization ability of $O\left(\sqrt{\frac{1}{m}}\right)$.

11.6.2 Regression from Bayesian perspective

Assume the following process:

- Choose w from normal distribution
- Given x , set $y = w \cdot x + R$ (where $R \sim N(\mu, 1)$ is noise).

Within the Bayesian approach we want to get the MAP and ML.

$$Pr[w|(x_1, y_1), ..., (x_n, y_n)] = \frac{Pr[(x_1, y_1), ..., (x_n, y_n)|w]Pr[w]}{Pr[(x_1, y_1), ..., (x_n, y_n)]}$$

$$= \left[\prod_{i=1}^{n} e^{-\frac{1}{2}(w \cdot x_i - y_i)^2}\right] e^{-\frac{\|w\|^2}{2\sigma^2}}$$

ML:

$$w_{ML} = \max_w Pr[(x_1, y_1), ..., (x_n, y_n)|w] = \min_w \frac{1}{2} \sum_{i=1}^{n} (w \cdot x_i - y_i)^2$$

This is actually the standard linear regression problem.

MAP:

$$w_{MAP} = \max_w Pr[(x_1, y_1), ..., (x_n, y_n)|w]Pr[w] = \min_w \|w\|^2 \frac{1}{2\sigma^2} + \frac{1}{2} \sum_{i=1}^{n} (w \cdot x_i - y_i)^2$$

Giving us the Ridge regression.

11.7 Summary

Model Selection deals with finding a good hypothesis based on a given number of examples.

We introduced the problem of overfitting our hypothesis to the particular examples. This problem arises only when we are limited in the number of examples, and therefore cannot require a large number of examples to sufficiently reduce the probability of error.

We presented two approaches for dealing with the overfitting problem: *Structural Risk Minimization* and *Cross Validation*.

Both of these methods deal with the overfitting problem by repressing the tendency towards overfitted hypotheses, which stems from the desire to minimize the hypothesis' error-level on the given data. *Structural Risk Minimization* does this by imposing a penalty on complexity; *Cross Validation* does it by learning the concept from only a portion of the given data, and then testing the hypothesis on the remaining portion of the data.

We showed that the *Cross Validation* approach approximates the results of the *Structural Risk Minimization*.

We then introduced a third method, *Minimum Description Length*, which strives to minimize the description-length of the chosen hypothesis. We showed that this method is basically a variation on the theme of *Structural Risk Minimization*. It values a hypothesis based on its observed error (amount of corrections which have to be described) and simplicity of the hypothesis (its description-length).

We ended by looking at regularization and linear regression. We saw that without regularization the solution to the regression can be highly instable. We saw two popular methods for regularizing regression.