Introduction to Machine Learning **Fall Semester, 2013**

Recitation 4: November 3

Lecturer: Mariano Schain *Scribe: ym*

4.1 Outline

The recitation reviewed first the lecture on Bayesian inference, discussing maximum likelihood, and EM for mixture of Gaussian. (The material is in the scribe of the previous recitation.)

The second part is concern with the PAC model, deriving a sample bound for rectangles.

4.2 Learning Rectangles

This part is identical to the lecture, up to the analysis, and is given for completeness

Suppose we want to predict what is a 'typical person'. Our input is a sample of different descriptions of people based on their height and weight and their label: whether it is a typical person ('+') or not ('-'). Let H be our hypothesis class. In our example, H will be the set of all possible rectangles on a two-dimensional plane which are axis-aligned (not rotated). One example of hypothesis $h \in H$ can be: mark a person which denotes as (height, weight) to be a typical one ('+') if its description is in the range of:

$$1.60 \leq height \leq 1.90, 60 \leq weight \leq 90$$

Assuming the real target function is also rectangle (This assumption will be inferred later on). Our goal is to find the best rectangle 'R'' that approximates the real rectangle target R. In general, we will try to learn an accurate predictor which will optimize our interests.

This learning model is different than the Bayesian inference where we assumed that the underlying distribution has a specific form and our goal is to estimate this distribution. In PAC, we don't know the underlying probability of the samples. Here, the typical people distribution is unknown and finding the common distribution of heights and weight is impossible without any knowledge about it.

Generally, in the PAC model, we won't impose any assumptions on the underlying distribution of the data other than that such a distribution exists and the samples are independently and identically distributed (i.i.d) according to the same distribution. If we wouldn't, and the samples were taken from different distribution than the data then we wouldn't have

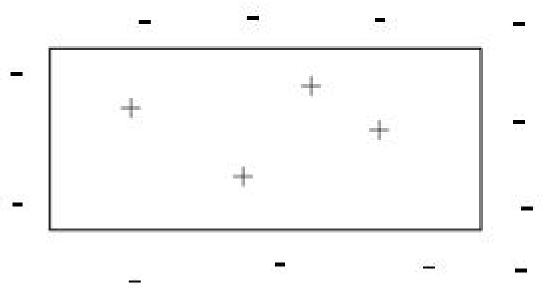

Figure 4.1: A rectangle with positive and negative examples

any reason to assume that we will be successful. This assumption gives us hope that what we learn based on the training set, gives good results that are close to the real target function.

The learning problem we described can be viewed as follows:

- Goal: Learn rectangle R.
- Input: Examples based on data set and their label: $\langle (x,y), +/- \rangle$.
- Output: R', a good approximation rectangle of the real target rectangle R. (An example of R', see Figure 4.1)

4.3 A Good Hypothesis

Our goal is to find a hypothesis that will have a small error rate, smaller than a defined ε. Let $R \Delta R'$ be the error of R' in respect to the real target rectangle R. This can be defined by two separate areas: $(R - R') \cup (R' - R)$ (where $(R - R')$ are false negative and $(R' - R)$ are false positive) as shown in Figure 4.2.

Then, our goal can be defined as to find R' with probability of at least $1 - \delta$ (confidence):

$$\Pr[error] = \mathcal{D}(R \Delta R') \leq \varepsilon$$

Assuming that $R \in H$ is the real target function.

4.4 Learning Strategy

Let $S = \{\langle (x_1, y_1), b_1 \rangle, ..., \langle (x_m, y_m), b_m \rangle\}$ be the sample data. Intuitive, we would like a rectangle that will be *consistent*, i.e., no error on the sample data. This is possible since we assumed that there exists a rectangle that labels correctly the data (the target rectangle R).

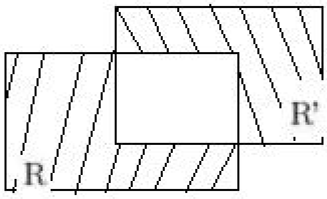

Figure 4.2: R and R' areas including two different error spaces. In our example, since $R' \subseteq R$ there exists only error of $(R - R')$.

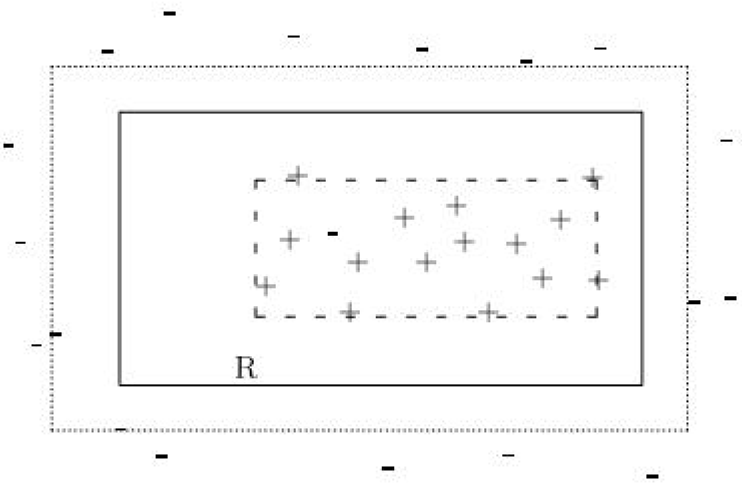

Figure 4.3: The smallest R_{min} (dashed line) and largest R_{max} (dotted line) rectangles border the rectangles which are consistent with the examples. The area between them represents the error region.

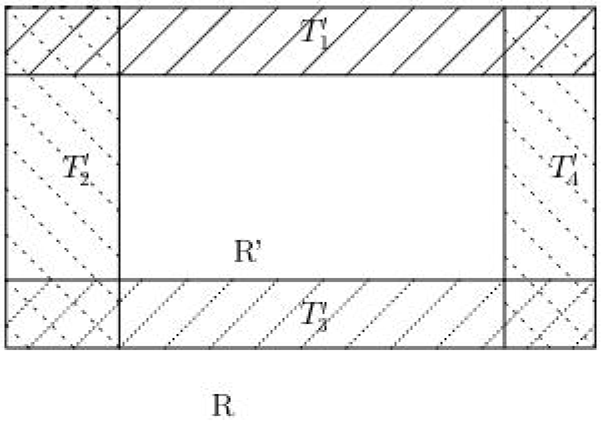

Figure 4.4: Breaking up the error into four rectangular strips $T'_1, ..., T'_4$

We can choose any rectangle varies from the minimal size (R_{min}) up to the maximize size (R_{max}) because it is *consistent* with the sample data, as shown in Figure 4.3. We will show later, that it doesn't matter which one of these rectangles the algorithm returns.

The strategy A we will try is to request a "sufficiently large" number of m examples, to choose the hypothesis rectangle R' that is the tightest fit to the positive examples ($R' = R_{min}$). That can be done by using four positive points indicating the left-most,right-most,down-most,up-most points in the sample data that are positive.

4.5 Sample Size

This is the main addition to what was presented in the lecture

In this section, we try to find what is a "sufficiently large" number of examples that is needed to learn a good hypothesis. For that, we will fix our accuracy and confidence parameters (ε, δ), and the strategy A. We will show that for any distribution D, we can assert sample size m that with high confidence (that is, probability at least $1 - \delta$), the returned rectangle R' from strategy A (i.e. the tightest fit rectangle) has an error of at most ε.

We will be aware of the fact that $R' \subseteq R$ and we construct $T'_1, .., T'_4$ areas that will indicate the error area, as defined in Figure 4.4. That is, $R \Delta R' = \cup T'_i$.

If the distribution of having each $T'_1, ..., T'_4$ is $D(T'_i) \leq \frac{1}{4}\varepsilon$, then the error rate of R' is at most: $\Pr[error] = \mathcal{D}(R \Delta R') = \mathcal{D}(\cup T'_i) \leq \sum_{i=1}^{4} \frac{\varepsilon}{4} = \varepsilon$ This analysis is incorrect because it depends on the strip T'_i that is constructed by the returned R' from the strategy A after seeing the sample. Because the sample has already been seen, the event $D(T'_i) < \frac{\varepsilon}{4}$ lost its randomness and in order to talk about it (its probability), we need to create an event which do not depend on the sample and use it in the proof.

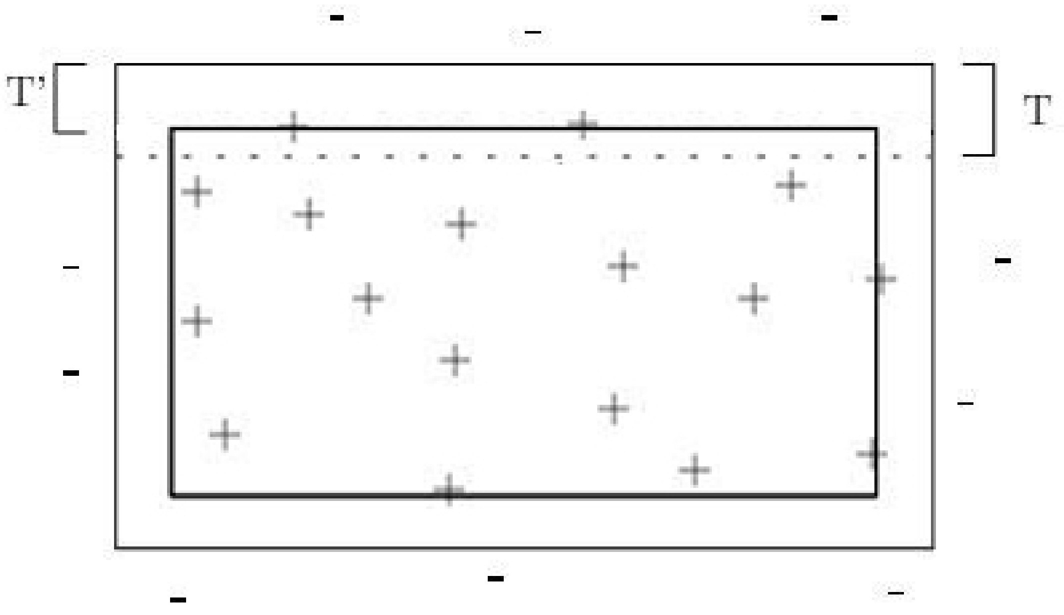

Figure 4.5: Adjusting strip size to have a weight of at most ε according to the real target function R. The strips $T_i' \subseteq T_i$ surrounding the inner rectangle $R' = R_{min}$ apart from the outer rectangle R.

A better construction will depend only on the real target rectangle R. We will construct strips $T_1, .., T_4$ such that $\forall i \; D(T_i) \leq \frac{1}{4}\varepsilon$ (Figure 4.5). From the construction we can see that T_i is independent of the sample and of R'. Note that we cannot certainly find the T_i but we can be sure that such T_i exists.

We would want to have $\forall i \; T_i' \subseteq T_i$. If that is the case, we obtained our requirement since

$$\Pr[error] = \mathcal{D}(R \Delta R') = \mathcal{D}(\cup T_i') \leq \mathcal{D}(\cup T_i) \leq \varepsilon$$

From the construction, if there is at least one sampled point that resides in T_i it implies that $T_i' \subseteq T_i$. This is true, since the rectangle from strategy A must include all sampled positive points in R. To achieve that we can ask: what is the probability of having a bad event, that is, what is the probability that we didn't receive points from the sample data that are located on the constructive strips, i.e., T_i. Formally,

$$\Pr[\text{error} > \varepsilon] = \Pr[\exists i = 1..4 \; \forall (x, y) \in S, (x, y) \notin T_i]$$

By definition of T_i,

$$\Pr[x \notin T_i] = 1 - \frac{\varepsilon}{4}$$

Since our sample data is i.i.d from distribution D:

$$\Pr[\forall(x,y) \in S, (x,y) \notin T_1] = (1 - \frac{\varepsilon}{4})^m$$

The same analysis holds on each T_i strip. Hence, we get that on the entire region the error would not exceed the sum of probabilities for each of the strips. That is,

$$\Pr[\text{error} > \varepsilon] \leq 4(1 - \frac{\varepsilon}{4})^m$$

From the inequality $(1 - x) \leq e^{-x}$, we obtain:

$$\Pr[\text{error} > \varepsilon] \leq 4(1 - \frac{\varepsilon}{4})^m \leq 4e^{-\frac{\varepsilon}{4}m} < \delta$$

That is, if we want to have accuracy ε and confidence of at least $1 - \delta$, we have to choose the sample size m to satisfy:

$$4e^{-\frac{\varepsilon}{4}m} < \delta <=> m > \frac{4}{\varepsilon} ln \frac{4}{\delta}$$

For this strategy A, and for every small (ε, δ) we like, we got the sample size that is needed for having a good learner.

4.5.1 Remarks

1. The analysis holds for any fixed probability distribution $\mathcal{D}$, we only required that the sample points are i.i.d from distribution $\mathcal{D}$ to obtain our bound.

2. The minimal sample size $m(\varepsilon, \delta)$ behaves as we might expect. One might want to have better accuracy by decreasing ε or greater confidence by decreasing δ — our algorithm requires more examples to meet those requirements. There is a stronger dependence in ε.

3. The parameter ε gives the degree of accuracy that we want to achieve. It determines what is a good hypothesis for achieving a good approximation in respect to the target function. In our example, the accuracy determines which of our hypothesis rectangles are good enough in respect to the real rectangle target. We pay attention that the accuracy does not depend on the data distribution.

4. The parameter δ gives the degree of confidence on having a good learner. Meaning, how sure are we that we've reached that level of accuracy. This can be related on, how typical the given sample data reflects the true distribution. Again, it does not depend on the data distribution.

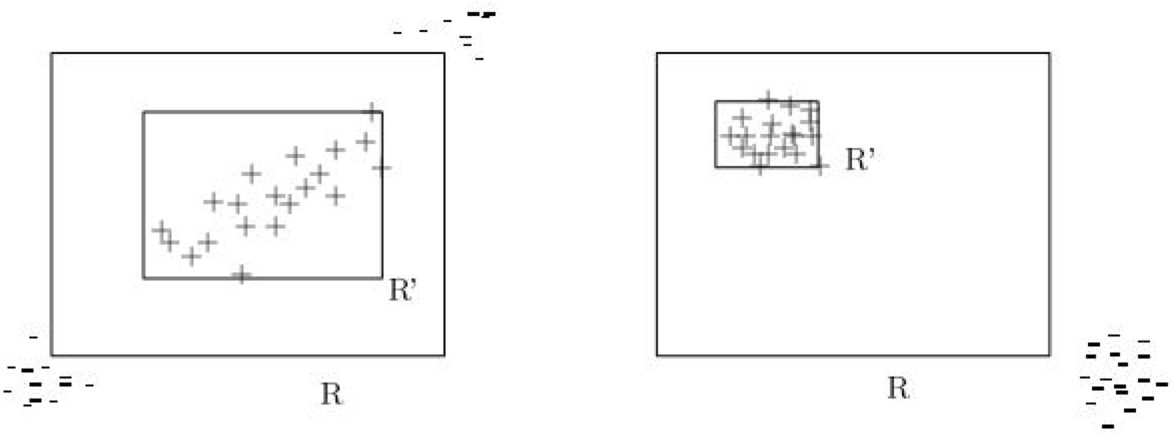

Figure 4.6: Two cases depending on the sample size

5. An example of learning, we might have cases as shown in Figure 4.6, where the distribution $\mathcal{D}$ gives large weights to particular regions of the plain, creating a distorted image of the rectangle. In any case, under those conditions, since the learner is tested on the same distribution $\mathcal{D}$, and this distribution has small error between R and R', the rectangle $\mathcal{R}'$ will be a good hypothesis (in respect to ε, δ).

6. The strategy A that we defined is efficient: In computational view, the only need is to search for the max and min points that defines our tightest-fit rectangle. In sample data size view, the number of examples that is required for achieving accuracy ε with confidence $1 - \delta$ is polynomial in $\frac{1}{\varepsilon}$ and $ln\frac{1}{\delta}$.

7. In this example, as opposed to the Bayesian approach, we haven't been trying to model $\mathcal{D}$ or to guess which rectangle is more likely (prior). We have separated the distribution $\mathcal{D}$ from the target function (rectangle R), and directly try to predict hypothesis for this function.