Chapter 9

Regression

The main difference between regression and pattern recognition is that the loss function $l(y, g(x))$ is real valued instead of being discrete. What we will do in this chapter is to outline how to modify the ideas for the pattern recognition problem to obtain generalization estimates for real valued loss functions, which includes regression.

It however turns out that dealing with unbounded loss functions is technically difficult and we will not cover it here, if you want more information, take a look at [SLT].

As in the pattern recognition problem, we want to bound

$$\mathbb{P}(\sup |R_n(\phi) - R(\phi)| > \epsilon) \tag{9.1}$$

we did that by rephrasing this as a problem of estimating empirical measures on certain classes of sets. How do we do the same for the problem where $l(y, g(x))$ can take any value between $[0, 1]$ for instance?

Consider a function $0 \leq \Phi(z) \leq 1$, and consider a sequence of i.i.d. random variables $Z, Z_i \sim \nu$, then

$$\begin{aligned}
\mathbb{E}\left[\phi(Z)\right] - \frac{1}{n}\sum_{i=1}^{n}\phi(Z_i) & \\
&= \int_0^1 \left(\nu(\Phi(z) > t) - \nu_n(\phi(z) > t)\right) dt \\
&\leq \sup_{\beta\in[0,1]} \left(\nu(\Phi(z) > \beta) - \nu_n(\phi(z) > \beta)\right) \int_0^1 dt \\
&= \sup_{\beta\in[0,1]} \left(\nu(\Phi(z) > \beta) - \nu_n(\phi(z) > \beta)\right)
\end{aligned}$$

where ν_n is the empirical measure based on $Z_1, \ldots, Z_n$.

That is, if we have a model space $\mathcal{M}$ of decision functions and a loss $l(y, g(x))$, for $g \in \mathcal{M}$ taking values in $[0, 1]$, then if we construct the class of

sets as follows

$$\mathcal{A} = \big\{\{(x,y) : l(y,g(x)) - \beta > 0\} : \beta \in (0,1)\big\}, \tag{9.2}$$

we can rewrite (9.1) in the following way

$$\mathbb{P}(\sup_{\mathcal{M}} |R_n(\phi) - R(\phi)| > \epsilon) \leq \mathbb{P}(\sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)| > \epsilon).$$

Now we are exactly the same situation as in (8.10) and we can apply Theorem 8.29 to get

Corollary 9.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let (X_1, Y_1), ..., (X_n, Y_n) be an i.i.d. sequence of random variables, X_i being continuous and taking values in $\mathbb{R}^m$ and $Y_i \in \mathbb{R}$. Then if $n\epsilon^2 \geq 8$ the following holds*

$$\mathbb{P}(\sup_{\mathcal{M}} |R_n(\phi) - R(\phi)| > \epsilon) \leq 8s(\mathcal{A}, n)e^{-n\epsilon^2/64}.$$

In the above, $\mathcal{A}$ is derived from $\mathcal{M}$ as in (9.2).

Example 9.2. *Assume that $g(x) \in [0,1]$ and that $l(y,x) = (y-x)^2$, then*

$$\{(x,y) : l(y,g(x)) - \beta > 0\} = \{(x,y) : |y - g(x)| > \sqrt{\beta}\}$$

that is, for every fixed value of x, the set is all y which are at distance greater than $\sqrt{\beta}$ from $g(x)$. That is, this is the complement to a tubular region around the graph $g(x)$.

A similar observation can be made for any convex loss function, it is thus clear that the growth function will often only depend on the complexity of $\mathcal{M}$ instead of depending on the choice of loss.

9.1 Guarantees with a held out testing set

Consider as in Section 8.3.1 the following notation. Consider a data set $T_{n+m} := \{(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})\}$ sampled i.i.d from $F_{X,Y}$. We consider $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ (which we dub the training data) and $\{(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})\}$ (which we dub the testing data). Define $\hat{\phi}$ the **empirical risk minimizer** on the **training dataset**, namely

$$\hat{R}_n(\hat{\phi}) = \min_{\phi \in \mathcal{M}} \hat{R}_n(\phi)$$

then since the **testing dataset** is independent of the training dataset and hence $\hat{\phi}$ is independent of the testing data, we want to use Theorem 3.6 but now we would like to consider the loss function to be unbounded. This

happens in the case of mean square error for instance. Consider the most usual quantity (mean squared error)

$$R(\phi) = \mathbb{E}\left[(Y - g(X))^2\right]$$

Previously we had $0-1$ loss and we thus knew that the random variable $L(Y, \phi(X))$ was bounded and we could immediately apply Theorem 3.6. Since we do not know that we have to make some assumptions to move forward, but this is getting into advanced topics and is outside the scope of this course, since how do we know that the assumptions make sense? Anyways, we will for simplicity make the assumption that $(Y - \phi(X))^2$ is sub-Gaussian with parameter $\lambda(\phi)$, i.e. where the parameter depends on the function ϕ. If you are up to it, you should spend some time thinking about why this is so. Hint: think of $\phi(x) = ax + b$ being a linear function and let $Y = 0$ and $X \sim \text{Bernoulli}(1/2)$, that is the sub-Gaussian parameter of $(\phi(X))^2$?

Anyways, we can get the following bound for the Risk using Theorem 3.13 that if $\hat{R}_m(\phi)$ denotes the empirical risk over the testing dataset we have

$$\mathbb{P}(|\hat{R}_m(\hat{\phi}) - R(\hat{\phi})| > \epsilon \mid T_n) < 2e^{-\frac{\epsilon^2 n}{2\lambda(\hat{\phi})^2}}. \tag{9.3}$$

We now run into some trouble, since we have a random variable on the right hand side of the bound, namely $\lambda(\hat{\phi})$ so we cannot use the tower property as we did in deriving (8.4) since we would need to be able to compute

$$\mathbb{E}\left[e^{-\frac{\epsilon^2 n}{2\lambda(\hat{\phi})^2}}\right].$$

However, this is usually not a problem. Since, if we adhere to the Train-Test philosophy we are actually only interested in (9.3), as at the point of having trained and found $\hat{\phi}$, the value $\lambda(\hat{\phi})$ is computable given only some assumptions on X.

Remark 9.3. *It should be noted that it is more often the case that* $(Y - \phi(X))^2$ *is sub-Exponential. This happens for instance if* Y *is Gaussian, since its squared, see Lemma 3.15.*

9.1.1 R^2

A common metric used in evaluating regression models is the so called R^2. The reason for the name comes from the theory of linear regression, where R^2 is actually the correlation squared. The metric is an empirical one. First consider what is called the fraction of variance unexplained

$$\widehat{FVU}(\hat{\phi}; T_{n+m} \setminus T_n) = \frac{\frac{1}{m}\sum_{i=n+1}^{m}(Y_i - \hat{\phi}(X_i))^2}{\frac{1}{m-1}\sum_{i=n+1}^{m}(Y_i - \frac{1}{m}\sum_{i=n+1}^{m} Y_i)^2} = \frac{\hat{R}_m(\hat{\phi})}{\hat{V}_m[Y]}$$

Lets explain the terms in the above expression as it looks quite crowded. The left hand side is the expression that we define as the Fraction of Variance Explained (FVU) and it takes the proposed (Trained) function $\hat{\phi}$ and the test set, i.e., $T_{n+m} \setminus T_n = \{(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})\}$. The ratio on the right hand side: the numerator is the empirical test risk $\hat{R}_m(\hat{\phi})$ and the denominator is the empirical variance of Y over the testing set.

Remark 9.4. *Due to the fact that in the definition of FVU the function $\hat{\phi}$ can be anything, we can have FVU take any non-negative value. Specifically, it can be greater than* 1.

Now, we define a version of R^2 that comes from the idea of variance explained and it is just $1 - FVU$, but beware, it can be negative so the term R^2 does not make sense, as it is not a square (only in the case of linear regression). You will encounter this confusion as you go out into industry, so make sure you get it right!

Can we make a concentration statement about FVU? Well, easiest is to realize that the true FVU given $\hat{\phi}$ is just

$$FVU(\hat{\phi}) = \frac{R(\hat{\phi})}{\mathbb{V}[Y]}$$

we can given the above discussion actually get intervals for both the numerator and denominator separately, i.e. if we assume that Y^2 is sub-Gaussian with parameter σ we get

$$\mathbb{P}(|\hat{R}_m(\hat{\phi}) - R(\hat{\phi})| > \epsilon \mid T_n) < 2e^{-\frac{\epsilon^2 n}{2\lambda(\hat{\phi})^2}}$$
$$\mathbb{P}\left(\left|\hat{V}_m[Y] - \mathbb{V}[Y]\right| > \epsilon \mid T_n\right) < 2e^{-\frac{\epsilon^2 n}{2\sigma^2}}$$

Using the union bound Lemma 1.12 we get that

$$\mathbb{P}\left(|\hat{R}_m(\hat{\phi}) - R(\hat{\phi})| \leq \epsilon \text{ and } \left|\hat{V}_m[Y] - \mathbb{V}[Y]\right| \leq \epsilon \mid T_n\right) \geq 1 - 2e^{-\frac{\epsilon^2 n}{2\lambda(\hat{\phi})^2}} - 2e^{-\frac{\epsilon^2 n}{2\sigma^2}}$$

Thus we can write a bound for the ratio by rearranging a bit and assuming that all the quantities are non-negative, i.e. $\hat{R}_m(\hat{\phi}) - \epsilon \geq 0$ and $\hat{V}_m[Y] - \epsilon \geq 0$.

$$\mathbb{P}\left(\frac{\hat{R}_m(\hat{\phi}) - \epsilon}{\hat{V}_m[Y] + \epsilon} \leq \frac{R(\hat{\phi})}{\mathbb{V}[Y]} \leq \frac{\hat{R}_m(\hat{\phi}) + \epsilon}{\hat{V}_m[Y] - \epsilon} \mid T_n\right) \geq 1 - 2e^{-\frac{\epsilon^2 n}{2\lambda(\hat{\phi})^2}} - 2e^{-\frac{\epsilon^2 n}{2\sigma^2}}$$

However, the problem is that often we do not know much about what is sub-Gaussian etc. but we might be in a situation where things are bounded and one could hope to apply Theorem 3.6. The problem is that in this case,

often the sum of square of the residual is quite small (good model) and perhaps also the variance of Y is small, in this case the Hoeffding inequality is too rough and we will employ a stronger inequality called **Bennett's inequality**. It looks quite complicated, but all the things are computable and we do it in the notebooks:

Theorem 9.5 (Bennett's inequality)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables with finite variance such that $\mathbb{P}(X_i \leq b) = 1$ with mean zero. Let and $\sigma^2 = \mathbb{V}[X_i]$. Then for any $\epsilon > 0$,*

$$\mathbb{P}(\overline{X}_n \geq \epsilon) \leq \exp\left(-\frac{n\sigma^2}{b^2} h\left(\frac{b\epsilon}{\sigma^2}\right)\right)$$

where $h(u) = (1+u)\log(1+u) - u$ for $u > 0$.

In what case can we apply this theorem. Lets make the assumption that $|Y| \leq 1$ (can be done via scaling of the data). Consider again that we found our model $\hat{\phi}$ by training on T_n and let $b = \max_X \hat{\phi} - \min_X \hat{\phi}$. The constant b can be a bit tricky to get, but we can guess its value by taking the max and the min over the data.

$$\mathbb{P}(|\hat{R}_m(\hat{\phi}) - R(\hat{\phi})| > \epsilon_1 \mid T_n) \leq \exp\left(-\frac{n\sigma^2}{b^2} h\left(\frac{b\epsilon_1}{\sigma^2}\right)\right)$$

where $\sigma^2 = \mathbb{V}[L(Y, \hat{\phi}(X))]$ (can also be estimated from data). For the Y part we can do

$$\mathbb{P}\left(\left|\hat{V}_m[Y] - \mathbb{V}[Y]\right| > \epsilon_2 \mid T_n\right) < \exp\left(-n\sigma^2 h\left(\frac{\epsilon_2}{\sigma^2}\right)\right)$$

where $\sigma^2 = \mathbb{V}[(Y - \mathbb{E}[Y])^2]$. One option is to find ϵ_1, ϵ_2 such that

$$\frac{\alpha}{2} = \exp\left(-\frac{n\sigma^2}{b^2} h\left(\frac{b\epsilon_1}{\sigma^2}\right)\right)$$
$$\frac{\alpha}{2} = \exp\left(-n\sigma^2 h\left(\frac{\epsilon_2}{\sigma^2}\right)\right)$$

which then gives a final bound of

$$\mathbb{P}\left(\frac{\hat{R}_m(\hat{\phi}) - \epsilon_1}{\hat{V}_m[Y] + \epsilon_2} \leq \frac{R(\hat{\phi})}{\mathbb{V}[Y]} \leq \frac{\hat{R}_m(\hat{\phi}) + \epsilon_1}{\hat{V}_m[Y] - \epsilon_2} \mid T_n\right) \geq 1 - \alpha$$

Remark 9.6. *There is an example in the Regression notebook where this inequality is used in practice.*

9.2 Bibliography

The generalization bounds and definition of the VC dimension for a more general loss than in the pattern recognition case is a very natural extension. Some parts of the above can be found in [SLT], but the connection to empirical measures was not done. As we know, we should expect some of these results to hold for unbounded loss functions, provided that we can bound the tail of the empirical risk. As far as I know, the most general results can be found in [SLT], Chapter 5. For Bennett's inequality, see for instance [BLM].