Machine Learning @ Data science 2019: Set 3

Jordi Morera Serra

March 15, 2019

Exercise 8

We define a circle as $C_{\boldsymbol{c},r} = \{\boldsymbol{x} \in \mathbb{R}^2 : \|\boldsymbol{x} - \boldsymbol{c}\| \leq r\}$.

Case 1 First we focus in the case where we have no restrictions regarding circle radius, $\mathcal{A} = \{C_{\boldsymbol{c},r} : \boldsymbol{c} \in \mathbb{R}^2, r \geq 0\}$.

Note that for number of points $n < 4$, $\mathcal{A}$ shatters all points. However, for $n \geq 4$,it is not possible to shatter them. Hence the *VC dimensions is 3.* This can be observed in figure 1. We cannot have a circle $\mathcal{A}'$ (red circle) that contains d and b but not c and a. This is a general situation for $n \geq 4$.

Case 2 Now we focus in the case where we restrict the radius $r = 1$,$\mathcal{A}_1 = \{C_{c,1} : c \in \mathbb{R}^2\}$.

If the points have pairwise distance higher than one, $\mathcal{A}_1$ cannot be shattered even with $n = 2$ because there is no $\mathcal{A}_1^{'}$ that contain both points. However, if the pairwise distance between our points is lower or equal than 1, then we find ourselves in the same situation than the previous case. In figure 1 we see that no $\mathcal{A}_1^{'}$ can contain d and b but not a and c. Hence, *the VC dimencions is also 3.*

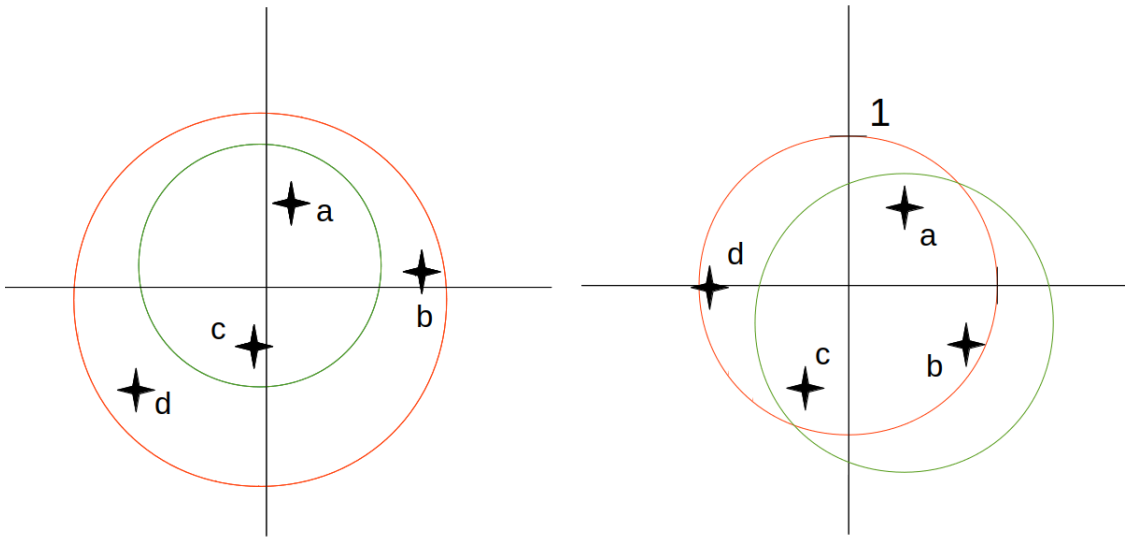


Figure 1: Case 1 and case 2 exemplification

Exercise 9

A half plane is a set of the form $\mathcal{H}_{a,b,c} = \{(x, y) \in \mathbb{R}^2 : ax + by \geq c\}$ for some real numbers a, b, c.

Case 1 If $c = 0$, we have a collection of half planes that cross the origin. The VC dimension, in this case, is equal to 2.

According to Sauer's lema, $S_{\mathcal{A}}(n) \leqslant (n+1)^{V_{\mathcal{A}}}$. Hence in our case we get that $S_{\mathcal{A},}(n) \leqslant (n+1)^2$. However, we can a set sharper bound. Imagine we have n points distributed in the first quadrant forming an arc with center in te origin. The number of different sets of points we can form is $2n$. Note that all other point layout will lead to a lower number of sets. Hence, $S_{\mathcal{A},}(n) = 2n$.

Case 2 If $c \in \mathbb{R}$, we have a collection of half planes with non restriction on the intercept. The VC dimension, in this case, is equal to 3.

According to Sauer's lema, $S_{\mathcal{A}}(n) \leqslant (n+1)^{V_{\mathcal{A}}}$. Hence in our case we get that $S_{\mathcal{A}}(n) \leqslant (n+1)^3$. However, we can set a sharper bound.

As we know the VC dimension is 3, we know that the $S_{\mathcal{A}}(n) < 2^n$ for any $n \geq 4$. Since we can only split our points with half spaces, we can only form $n(n+1)$ groups of 1 to $n-1$ points. Then, adding the empty space and the group of all points, we finally get $S_{\mathcal{A}}(n) = n(n-1) + 2$.

In figure 2 we can see an example with $n = 4$. With a class set of half-spaces, I cannot group points d and a together and c and b together. The only grouping of two points I am able to do are $c$and d, d and b, etc. Therefore, for $n = 4$, I will only be able to group two points together in 4 different ways. If we add this result to the number of possible groups of 1 point, $\binom{4}{1}$, the empty space, $\binom{4}{0}$, and its complements, $\binom{4}{3}$ and $\binom{4}{4}$, we finally get $S_{\mathcal{A}}(4) \leqslant 14$.

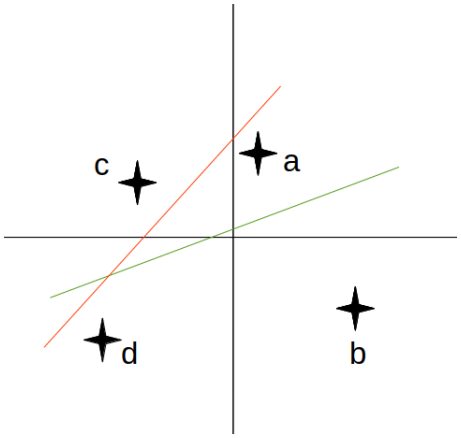


Figure 2: Forming groups of 2 points with half-spaces

Exercise 10

- Proof $\mathcal{R}_n(\mathcal{A} \cup \mathcal{B}) = \mathcal{R}_n(\mathcal{A}) + \mathcal{R}_n(\mathcal{B})$

Since$\mathcal{R}_n$ cannot be negative, we get,

$$\begin{aligned}
\mathcal{R}_n(\mathcal{A} \cup \mathcal{B}) &= \mathbb{E} \sup_{v \in \mathcal{A} \cup \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i v_i \right| \\
&= \mathbb{E} \sup_{a \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \right| + \mathbb{E} \sup_{b \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i b_i \right| - \mathbb{E} \sup_{\mathcal{A} \cap \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i c_i \right| \\
&\leq \mathbb{E} \sup_{a \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \right| + \mathbb{E} \sup_{b \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i b_i \right| \\
&= \mathcal{R}_n(\mathcal{A}) + \mathcal{R}_n(\mathcal{B})
\end{aligned}$$

- Proof $\mathcal{R}_n(c \cdot \mathcal{A}) = |c| \cdot \mathcal{R}_n(\mathcal{A})$

Taking Rademacher average definition,

$$\mathcal{R}_n(c \cdot \mathcal{A}) = \mathbb{E} \sup_{a \in \mathcal{A}} \left| c \cdot \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \right|$$

as c is a constant, we can take it out from the expectation and scale the supremum a posteriori.

$$\begin{aligned}
\mathbb{E} \sup_{a \in \mathcal{A}} \left| c \cdot \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \right| &= |c| \cdot \mathbb{E} \sup_{a \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \right| \\
&= |c| \cdot \mathcal{R}_n(\mathcal{A})
\end{aligned}$$

- Proof $\mathcal{R}_n(\mathcal{A} \oplus \mathcal{B}) \leq \mathcal{R}_n(\mathcal{A}) + \mathcal{R}_n(\mathcal{B})$

$$\begin{aligned}\mathcal{R}_n(\mathcal{A}+\mathcal{B}) &= \mathbb{E} \sup_{v \in \mathcal{A}+\mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i v_i \right| \\ &= \mathbb{E} \sup_{a \in \mathcal{A}, b \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(a_i + b_i\right) \right|\end{aligned}$$

Applying the triangle inequality, we finally get

$$\begin{aligned}\mathbb{E} \sup_{a \in \mathcal{A}, b \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(a_i + b_i\right) \right| &\leq \mathbb{E} \sup_{a \in \mathcal{A}, b \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \right| + \mathbb{E} \sup_{b \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i b_i \right| \\ &= \mathcal{R}_n(\mathcal{A}) + \mathcal{R}_n(\mathcal{B})\end{aligned}$$

- Proof $\mathcal{R}_n(\mathcal{A}) = \mathcal{R}_n(absconv(\mathcal{A}))$

Firstly, since $\mathcal{A} \subset absconv(\mathcal{A})$,then $\mathcal{R}_n(\mathcal{A}) \leq \mathcal{R}_n(absconv(\mathcal{A}))$. Also, applying the definition given of $absconv(\mathcal{A})$ and previous proofs, we get

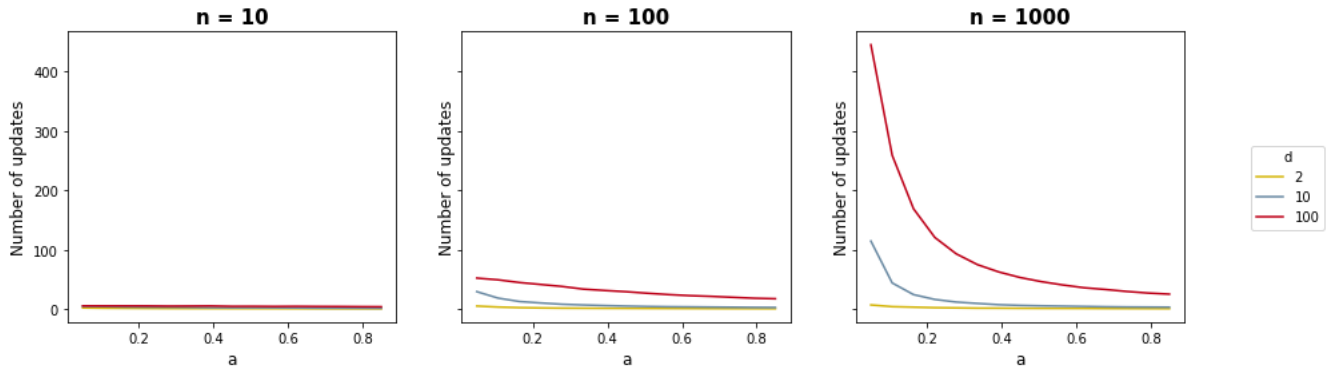
$$\mathcal{R}_n\left(c_1\mathcal{A} + \ldots + c_N\mathcal{A}\right) \leq \sum_{i=1}^{N} |c_i| \, \mathcal{R}_n(\mathcal{A}) \leq \mathcal{R}_n(\mathcal{A})$$

Since $absconv(\mathcal{A})$ is the union of all sets of the form $\left\{\sum_{j=1}^{N} c_j a^{(j)} : N \in \mathbb{N}, \sum_{j=1}^{N} |c_j| \leq 1, a^{(j)} \in A\right\}$, we see that $\mathcal{R}_n(absconv(\mathcal{A})) \leq \mathcal{R}_n(\mathcal{A})$. Therefore, $\mathcal{R}_n(absconv(\mathcal{A})) = \mathcal{R}_n(\mathcal{A})$.

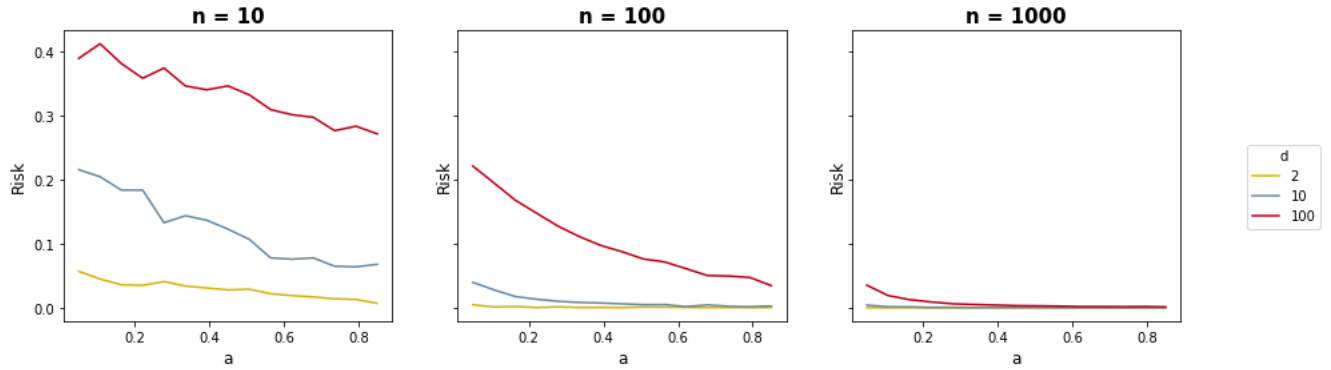
Exercise 11

Linearly separable data

In figure 3 we can see the number of updates and the estimated risk for different margin a values, d dimensions and n sample size for linearly separable data.



(a) Number of Perceptron algorithm updates



(b) Probability of error in different situations

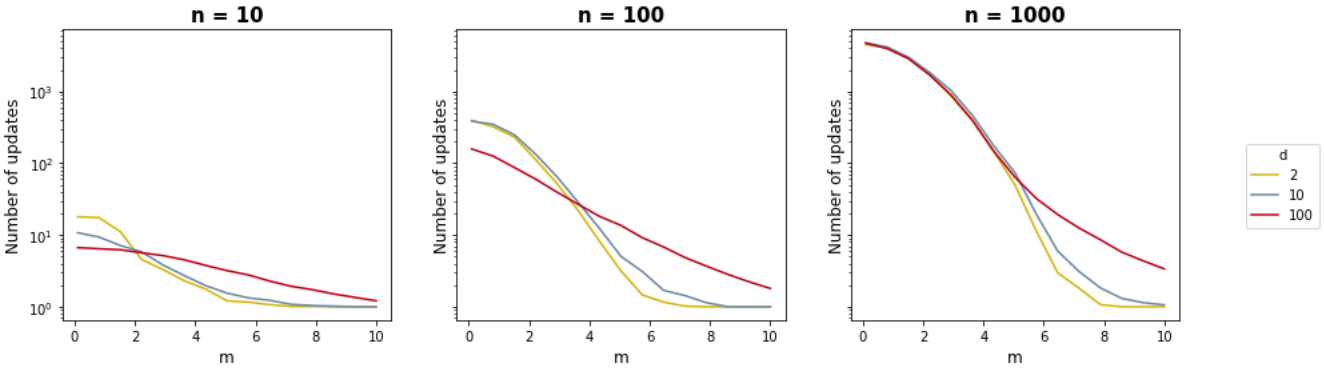
Figure 3

Remarks

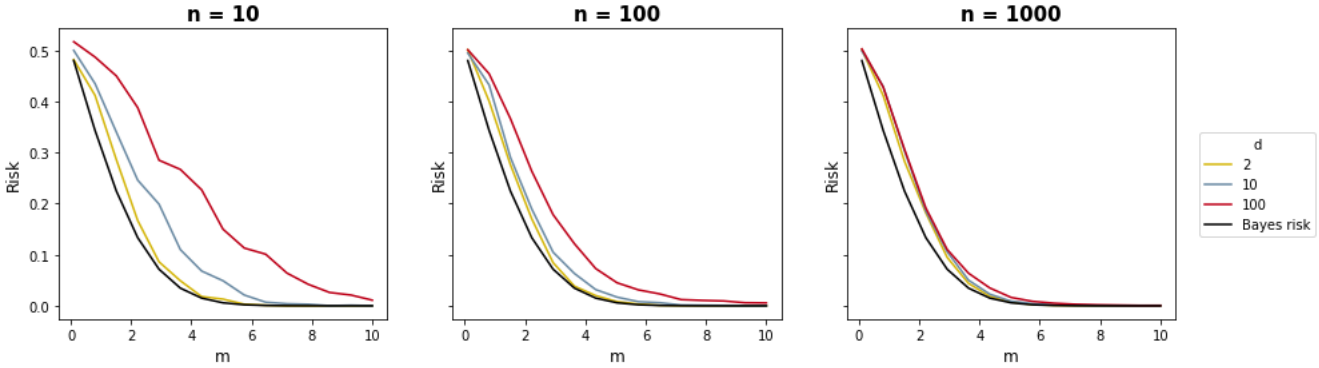
- According to Perceptron convergence theorem, when the data is linearly separable, the number of updates the algorithm does before converging is bounded by $\left(\frac{R^2}{a^2}\right)$, where $R = \underset{x_i}{max}||X_i||$ and a is the margin of the data. The euclidean norm of a vector will increase with the dimension, therefore, for a given, margin a the number of updates the algorithm has to do to stop are higher.
- When we increase the sample size, the probability of having an extreme point is higher, therefore, the probability of having a big R is higher. If we refer again to the Perceptron convergence theorem, we get that the number of updates bound is higher. This can be observed in figure 3-a.
- Regarding the risk, we clearly see it decreasing with the margin. This is what we expected since as the margin increases, the two conditional distributions are further apart and the classifier fitted with the training data is more likely to predict correctly out-training-sample data. Also note that when a is high enough it converges to the bayes risk, which is 0 for linearly separable data.
- When n increases, since a bigger training sample is available, the risk decreases no matter d. On the other hand, for a given n, having higher d adds noise to our classification problem since all dimensions but the first one are not relevant.

Not linearly separable data

In figure 4 we can see the number of updates and the estimated risk for different m values, d dimensions and n sample size for non linearly separable data. In this case, the algorithm never stops and stopping conditions have to be set. The algorithm will stop, either when the risk after having done one iteration over the training set does not change compared to the previous one, or, when we went over the whole training sample 10 times without converging on the risk.



(a) Number of Perceptron algorithm updates



(b) Probability of error in different situations

Figure 4

Remarks

- Notice that in figure 4 - a, the y scale is logarithmic. For m small, the algorithm iterates over the data until one of the stopping conditions is reached. If we analyze them, most of the times, the algorithm has stopped because the risk has not changed much from one iteration to the other. For these values of m, the risk is really high. This means that, altough the algorithm stopped and we obtained a vector of weights w, there are several w' that could reach this level of risk. Therefore, the number of updates obtained is not relevant. However, as m increases and the data becomes "more" linearly separable, we found ourselves with similar situation as the previous case where high dimensional classification problems require more updates to converge.
- As in the linearly separable case, as n increases we increase the number of updates.
- Since data becomes linearly separable as m increases, the risk tends to 0 when m increases for all n and d.
- Per construction of the problem, the bayes risk is the same no matter d. As expected, is always below for all different linear classifiers. For similar reasons than the linearly separable case, for a given n, as d increases, the classification problem becomes noisier and, hence, the risk is higher.