10-601 Machine Learning: Homework 5

Due 5 p.m. Wednesday, February 25, 2015

Instructions

- **Late homework policy:** Homework is worth full credit if submitted before the due date, half credit during the next 48 hours, and zero credit after that. You *must* turn in at least $n-1$ of the n homeworks to pass the class, even if for zero credit.
- **Collaboration policy:** Homeworks must be done individually, except where otherwise noted in the assignments. "Individually" means each student must hand in their own answers, and each student must write and use their own code in the programming parts of the assignment. You may collaborate with others on this problem set and consult external sources. ***However, you must write your own solutions and fully list your collaborators/external references for each problem.*** We will be assuming that, as participants in a graduate course, you will be taking the responsibility to make sure you personally understand the solution to any work arising from such collaboration.
- **Online submission:** You must submit your solutions online on autolab. We recommend that you use LaTeX to type your solutions to the written questions, but we will accept scanned solutions as well. On the Homework 4 autolab page, you can download the template, which is a tar archive containing a blank placeholder pdf for the written questions and one Octave source file for each of the programming questions. Replace each pdf file with one that contains your solutions to the written questions and fill in each of the Octave source files with your code. When you are ready to submit, create a new tar archive of the top-level directory and submit your archived solutions online by clicking the "Submit File" button. You should submit a single tar archive identical to the template, except with each of the Octave source files filled in and with the blank pdf replaced by your solutions for the written questions. You are free to submit as many times as you like (which is useful since you can see the autograder feedback immediately).

  DO NOT change the name of any of the files or folders in the submission template. In other words, your submitted files should have exactly the same names as those in the submission template. Do not modify the directory structure.

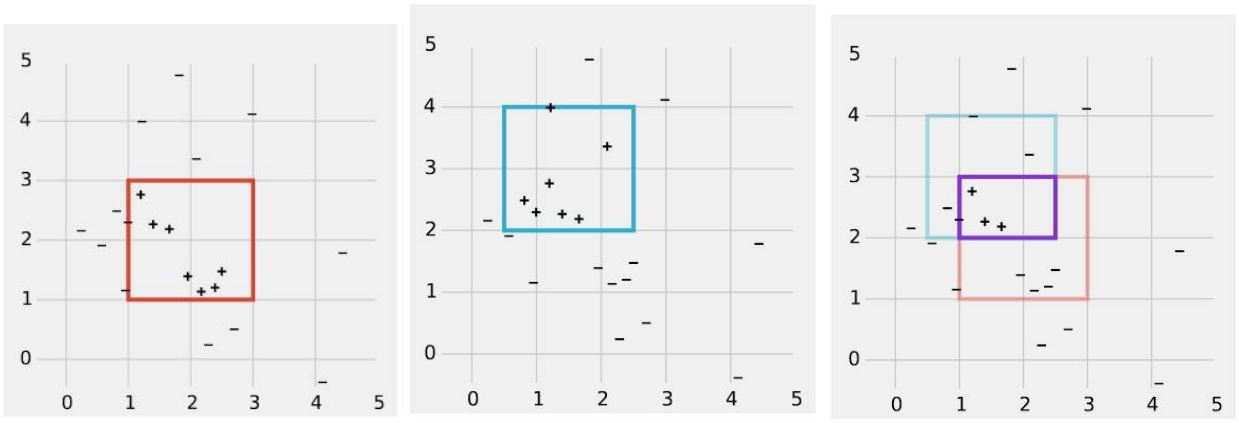
Problem 1: VC Dimension

Recall that we call a set of points *shattered* by a class of functions H if all possible $\{-1, +1\}$ labelings of the points can be produced by some function in H. The *Vapnik-Chervonenkis* (VC) dimension is the size of the largest set of points that can be shattered by the hypothesis space. See the lecture notes, video, and recitation video for more information.

In this problem, we will explore the hypothesis space where each hypothesis is a combination of two simpler hypotheses. More precisely, given two hypotheses h_1 and h_2, we define $h = h_1 \cap h_2$ as a new hypothesis that labels an example $+1$ only if both h_1 and h_2 give the label $+1$, otherwise, it is labeled -1. We can extend this to *sets* of hypotheses: given two sets of hypotheses H_1 and H_2, define $H^* = \{h_1 \cap h_2 : h_1 \in H_1, h_2 \in H_2\}$ as the set of all intersections of hypothesis pairs from the two classes H_1 and H_2.

As an example, let H_1 be the set of classifiers in $\mathbb{R}$ that assigns the label $+1$ if the example is larger than some threshold a. Let H_2 be the set of classifiers in $\mathbb{R}$ that assigns the label $+1$ if the example is smaller than some threshold b. Then H_* would be the set of all intervals (a, b) in $\mathbb{R}$ that assigns $+1$ if the example is inside the interval. Another example is when H_1 and H_2 is the set of all (axis-aligned) squares in $\mathbb{R}^2$, H^* is

the set of all axis-aligned rectangles. This example is illustrated below. On the left, we have a single square classifier h_1; in the middle we again have a square classifier h_2; and on the right, we have $h_1 \cap h_2$, which is a rectangle classifier.



Keep in mind that these are only examples. We are looking for results that can apply generally to any pair of hypotheses classes.

a. **[15 pts]** Suppose that the *shattering coefficient* of H_1 is $H_1[n]$ (i.e. the maximum number of ways that the hypothesis space H_1 can label a set of n points is $H_1[n]$). Similarly, suppose that the shattering coefficient of H_2 is $H_2[n]$. Show that $H^*[n] \leq H_1[n]H_2[n]$.

Solution

We know that $|H^*| \leq |H_1||H_2|$. (kev:) does this still work for infinite H?
In the upper bound, where $|H^*| = |H_1||H_2|$ and H^* has no repeated elements, then we can see that

$$H^*[n] \leq H_1[n]H_2[n]$$

b. **[5 pts]** Show that if the VC dimension of a hypothesis space H is d, then the shattering coefficient $H[d]$ is 2^d.

Solution

A result of Sauer's lemma is that if $m \leq VCdim(H)$, then $H[m] = 2^m$.
Proof here
Then it follows that $H[d] = 2^d$.

c. **[15 pts]** Let H be a hypothesis space with VC dimension d. Define H^* as the hypothesis space produced by all intersections of pairs of hypotheses from H (assuming that $H_1 = H_2 = H$ in our above definitions). Use your results from **a.** and **b.** to show that the VC dimension d_* of H^* is bounded by $\mathcal{O}(d \log d)$. You may use the fact that if $2^x \leq x^y$, then $x \leq k \cdot y \log y$ for some constant k.

Hint: Since d_* is the VC dimension of H^*, then by definition, there exists a set S of d_* points that is shattered by H^*. By Sauer's lemma, we know that the maximum number of ways that H can label S is bounded by $\mathcal{O}(d_*^d)$. That is, $H[d_*] = \mathcal{O}(d_*^d)$.

Solution

From **a.** we can say that:

$$H^*[d_*] \leq (H[d_*])^2$$

Using the hint, the right side becomes:

$$H^*[d_*] = \mathcal{O}(d_*^{2d})$$

Using **b.**, the left side becomes:

$$2^{d_*} = \mathcal{O}(d_*^{2d})$$

Now we pick an arbitrary constant α s.t.

$$2^{d_*} = \mathcal{O}(d_*^{2d}) \leq d_*^{\alpha d}$$

Using the inequality given, that if $2^x \leq x^y$, then $x \leq k \cdot y \log y$,

$$d_* \leq k \cdot \alpha d \log(\alpha d)$$

Simplifying the expression,

$$d_* = \mathcal{O}(d \log d)$$

d. For each one of the following function classes, find the VC dimension. State your reasoning.

i. **[4 pts]** Half spaces in $\mathbb{R}$, where examples on one side of the boundary are labeled $+1$, and examples on the other side are labeled -1.

Solution

You can easily shatter a set of 1 point by placing the boundary on both sides of the point. You cannot shatter a set of 2 points because there are only 3 places to put the boundary: left of both points, middle, and right of both points. Therefore the VC dimension is 1.

ii. **[4 pts]** Half spaces in $\mathbb{R}^2$, where examples on one side of the line are labeled $+1$, and examples on the other side are labeled -1.

Solution

The VC dimension is 3 because we know that the VC dimension of linear separators in $\mathbb{R}^d$ is $d + 1$. We can shatter a set of 3 points arranged in a triangle. We cannot shatter a set of 4 pts where 1 pt is in the convex hull.

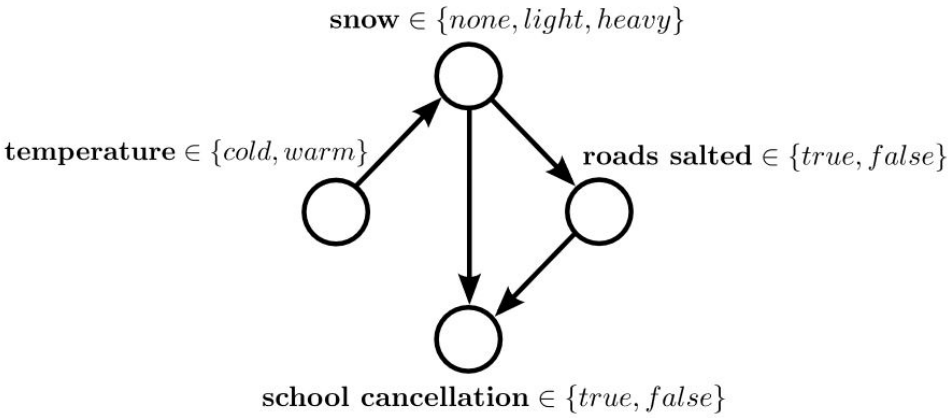
iii. **[7 pts]** Axis-aligned squares in $\mathbb{R}^2$, where points are labeled $+1$ inside the square, and -1 outside (as in the illustrations above).

Solution

The VC dimension is 4. We can shatter 4 pts arranged in a diamond shape. We cannot shatter a set S of 5 pts. Consider a subset G of size at most 4, containing the points with max x, min x, max y, and min y. Any axis-aligned rectangle that contains G must also contain all the pts. Now if we draw a rectangle around G, we notice that the remaining points must lie within the rectangle as well. Therefore, the labeling where G is +1 and the other points are -1 is impossible.
See here for reference

Problem 2: Graphical Models

Below is depicted a graphical model with four *discrete* random variables that can be used to predict whether school will be closed due to inclement weather. **Note:** We will cover concepts that will help in completing this problem in recitation on Feb 19th and in lecture on Feb 23rd. The videos will be posted on the website.



a. Answer the following questions about the conditional independence structure in the model:

i. **[4 pts]** Which variables are independent of **temperature** given that **snow** is observed?

Solution

school cancellation and **roads salted**

ii. **[4 pts]** Which variables are independent of **snow** given that no variables are observed?

Solution

There are none.

iii. **[4 pts]** Which variables are independent of **snow** given that **temperature** is observed?

Solution

There are none.

iv. **[4 pts]** Which variables are independent of **school cancellation** given that **snow** and **roads salted** are observed?

Solution

temperature

Suppose the random variables in the above graphical model have the following parameters:

The variable **temperature** does not depend on any other variable, and so it has the following prior distribution:

$p(\textbf{temperature} = cold)$	$p(\textbf{temperature} = warm)$
0.4	0.6

The variable **snow** only depends on the value of **temperature**:

temperature	$p(\textbf{snow} = \textit{none} \mid \textbf{temp})$	$p(\textbf{snow} = \textit{light} \mid \textbf{temp})$	$p(\textbf{snow} = \textit{heavy} \mid \textbf{temp})$
cold	0.4	0.4	0.2
warm	0.9	0.08	0.02

The variable **roads salted** only depends on the value of **snow**:

snow	$p(\textbf{roads salted} = T \mid \textbf{snow})$	$p(\textbf{roads salted} = F \mid \textbf{snow})$
none	0.01	0.99
light	0.9	0.1
heavy	0.97	0.03

The variable **school cancellation** depends on both **snow** and **roads salted**. For brevity, the condition "**snow**, **roads slated**" is replaced with "...":

snow	**roads salted**	$p(\textbf{school cancellation} = T \mid \ldots)$	$p(\textbf{school cancellation} = F \mid \ldots)$
none	T	0.01	0.99
none	F	0.01	0.99
light	T	0.2	0.8
light	F	0.4	0.6
heavy	T	0.95	0.05
heavy	F	0.99	0.01

b. **[4 pts]** The joint probability is given by $p(\textbf{temperature}, \textbf{snow}, \textbf{roads salted}, \textbf{school cancellation})$. Write the factorized form of the joint probability (as a product of simpler probabilities) for the model above.

Solution

We use the following abbreviations: **temperature** $\to$ **T**, **snow** $\to$ **S**, **roads salted** $\to$ **RS**, **school cancellation** $\to$ **SC**.

$$p(\mathbf{T}, \mathbf{S}, \mathbf{RS}, \mathbf{SC}) = p(\mathbf{T})p(\mathbf{S} \mid \mathbf{T})p(\mathbf{SC} \mid \mathbf{S}, \mathbf{RS})p(\mathbf{RS} \mid \mathbf{S})$$

c. Using the above model, compute the following quantities. Show your work.

i. **[6 pts]** What is the probability

$$p(\textbf{temperature} = \textit{cold},\ \textbf{snow} = \textit{light},\ \textbf{roads salted} = F,\ \textbf{school cancellation} = T)?$$

Solution

Using the factorized form,

$$p(\mathbf{T} = \textit{cold})p(\mathbf{S} = \textit{light} \mid \mathbf{T} = \textit{cold})p(\mathbf{SC} = T \mid \mathbf{S} = \textit{light}, \mathbf{RS} = F)p(\mathbf{RS} = F \mid \mathbf{S} = \textit{light})$$

$$= 0.4 \cdot 0.4 \cdot 0.4 \cdot 0.1 = 0.0064$$

ii. **[6 pts]** Compute the distribution $p(\textbf{snow} \mid \textbf{school cancellation} = T, \textbf{temperature} = \textit{cold})$.

Solution

Using the definition of conditional probability,

$$p(S = none \mid SC = T, T = cold) = \frac{p(S = none, SC = T, T = cold)}{p(SC = T, T = cold)}$$

Now we express the numerator and denominator as marginalized forms of the full joint distribution.
For the numerator:

$$num = p(S = none, SC = T, T = cold) = \sum_{rs} p(S = none, SC = T, T = cold, RS = rs)$$

$$= p(T = cold)p(S = none|T = cold) \sum_{rs} p(SC = T \mid S = none, RS = rs)p(RS = rs \mid S = none)$$

$$= 0.4^2 \left[0.01 \cdot 0.99 + 0.01^2\right]$$

Likewise we can calculate the numerator for the other values of **snow**, then normalize by the sum of all the numerators.

iii. **[6 pts]** Compute the distribution $p(\textbf{snow} \mid \textbf{school cancellation} = F, \textbf{temperature} = cold)$.

Solution

This question follows the same process as **2 c ii**, except we set SC = F.

iv. **[6 pts]** Compute the distribution $p(\textbf{school cancellation} \mid \textbf{snow} = light)$.

Solution

$$p(SC \mid S = light) = \frac{\sum_{rs} p(SC, S = light, RS = rs)}{p(S = light)}$$

$$= \frac{\sum_{rs} p(SC \mid S = light, RS = rs)p(RS = rs \mid S = light)p(S = light))}{p(S = light)}$$

$$= \sum_{rs} p(SC \mid S = light, RS = rs)p(RS = rs \mid S = light)$$

Now we can just read values off of the conditional probability table to calculate the probability distribution.

v. **[6 pts]** What is the probability

$$p(\textbf{school cancellation} = T \mid \textbf{temperature} = cold,\ \textbf{snow} = light,\ \textbf{roads salted} = F)?$$

Solution

school cancellation is conditionally independent of **temperature** given the other 2 observations, so we can remove it:

$$p(\textbf{school cancellation} = T \mid\ \textbf{snow} = light,\ \textbf{roads salted} = F)$$

Now we just read out the table value, which is 0.4.

Problem 3: Extra Credit

In this optional extra credit problem, we will derive the VC dimension of half-spaces in $\mathbb{R}^n$. Note that all linear classifiers fall into this class (logistic regression, perceptrons, support vector machines, etc). Let H_n be the set of half-spaces in $\mathbb{R}^n$.

a. [5 pts] Lower bound. Prove that VC-dim$(H_n) \geq n + 1$ by presenting a set of $n + 1$ points in n-dimensional space such that one can partition that set with halfspaces in all possible ways. (And, show how one can partition the set in any desired way.)

Solution

We can place a point at the origin and n points with a 1 in one coordinate and zeros in the rest. We can partition the set arbitrarily into two subsets S_1 and S_2 with a hyperplane (let the origin be in S_1). Then just choose the hyperplane:

$$\sum_{\{i:p_i \in S_2\}} x_i = 1/2$$

(kev:) ? Reference

b. [5 pts] Upper bound. The following is "Radon's Theorem," from the 1920's.

Theorem. *Let S be a set of $n + 2$ points in n dimensions. Then S can be partitioned into two (disjoint) subsets S_1 and S_2 whose convex hulls intersect.*

Show that Radon's Theorem implies that the VC-dimension of halfspaces is *at most* $n + 1$. Conclude that VC-dim$(H_n) = n + 1$.

Solution

Consider the halfspace s.t. S_1 is assigned +1, and S_2 is labeled -1. Note that if a halfspace assigns a particular label to a set of points, then every point in their convex hull is also assigned the same label. However the intersection of the convex hulls is non-empty, so there is a contradiction when we try to label those points.

c. [5 pts] Now we prove Radon's Theorem. We will need the following standard fact from linear algebra. If $x_1, \ldots, x_{n+1}$ are $n + 1$ points in n-dimensional space, then they are linearly dependent. That is, there exist real values $\lambda_1, \ldots, \lambda_{n+1}$ *not all zero* such that $\lambda_1 x_1 + \ldots + \lambda_{n+1} x_{n+1} = 0$.

You may now prove Radon's Theorem however you wish. However, as a suggested first step, prove the following. For any set of $n+2$ points $x_1, \ldots, x_{n+2}$ in n-dimensional space, there exist $\lambda_1, \ldots, \lambda_{n+2}$ *not all zero* such that $\sum_i \lambda_i x_i = 0$ and $\sum_i \lambda_i = 0$. (This is called *affine dependence*.)

Solution

(kev:) todo ref1 , ref2