Discussion 7: VC Dimension
Probabilistic Machine Learning, Spring 2018

1 Hoeffding's Inequality

a) Chernoff Bounds: Let X be a random variable, prove that, for any $t \geq 0$

$$\Pr(X \geq \mu_X + t) \leq \min_{\lambda \geq 0} \mathbb{E}[e^{\lambda(X-\mu_X)}]e^{-\lambda t},$$

where $\mu_X = \mathbb{E}[X]$ is the mean of X.

Solution:

$$\begin{aligned}
\Pr(X \geq \mu_X + t) &= \Pr(X - \mu_X \geq t) \\
&\leq \Pr(e^{(X-\mu_X)\lambda} \geq e^{\lambda t}) && \text{(True for all } \lambda \geq 0) \\
&\leq \frac{\mathbb{E}[e^{(X-\mu_X)\lambda}]}{e^{\lambda t}} && \text{(Markov's inequality)} \\
&= \mathbb{E}[e^{(X-\mu_X)\lambda}]e^{-\lambda t} \\
&= \min_{\lambda \geq 0} \mathbb{E}[e^{(X-\mu_X)\lambda}]e^{-\lambda t}. && \text{(Statement still holds for the min because it's true for all } \lambda)
\end{aligned}$$

b) Hoeffding's Lemma: Let X be a bounded random variable with $X \in [a, b]$. Then

$$\mathbb{E}[e^{\lambda(X-\mu_X)}] \leq \exp(\frac{\lambda^2(b-a)^2}{8}), \text{ for all } \lambda \in \mathbb{R}.$$

Use Chernoff bounds and Hoeffding's lemma to prove Hoeffding's inequality

$$Pr(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_{X_i}) \geq t) \leq \exp(-\frac{2nt^2}{(b-a)^2}), \text{ for all } t \geq 0.$$

where $X_1, ..., X_n$ are independent random variables with $X_i \in [a, b]$ for all i.

Solution:

$$\begin{aligned} Pr(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \geq t) &= Pr(\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \geq nt) \\ &\leq \mathbb{E}[\exp(\lambda \sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]))]e^{-\lambda nt} \\ &= (\prod_{i=1}^{n}\mathbb{E}[e^{\lambda(Z_i - \mathbb{E}[Z_i])}])e^{-\lambda nt} \\ &\leq (\prod_{i=1}^{n} e^{\frac{\lambda^2(b-a)^2}{8}})e^{-\lambda nt} \end{aligned}$$

Rewriting this slightly and minimizing over λ, we have

$$Pr(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \geq t) \leq \min_{\lambda \geq 0} \exp(\frac{n\lambda^2(b-a)^2}{8} - \lambda nt) \quad (1)$$

$$= \exp(-\frac{2nt^2}{(b-a)^2}). \quad (2)$$

c) Hoeffding's inequality is very loose in certain cases. Please give a simple distribution of X_i where the bound can be much sharper than Hoeffding's bound.

Solution: Since Hoeffding's inequality only depends on the upper and lower bounds a and b of X_i, it can be very loose when the X_i has low variance. Any example with X_i having a low variance should be good.

For example, compare (i) $X_i = -1$ or $X_i = +1$, each with probability 0.5; and (ii) $X_i = 0$ with probability 0.98 and $X_i = -1$ or $X_i = +1$, each with probability 0.01. Example (ii) should intuitively enjoy a sharper bound because it has smaller variance.

2 Vapnik-Chervonenkis (VC) Dimension

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

1. Closed intervals in $\mathbb{R}$: $f : \mathbb{R} -> \{0, 1\}$, where an example is labeled positive if it lies within the interval, and negative otherwise:

   $$H = \{f(x) = I_{x \in [a,b]}\}$$

   Solution VCdim = 2. We can always shatter 2 points on a line, since there is only a single block of positive points that could be captured by the same interval. However 3 points can not be shattered due to the arrangements such as positive-negative-positive.

2. Union of 2 intervals in $\mathbb{R}$: $f : \mathbb{R} -> \{0, 1\}$, where an example is labeled positive if it lies inside one of the intervals, and negative otherwise.

   $$H = \{f(x) = I_{x \in [a,b] \cup [c,d]}\}$$

   Solution VCdim = 4. A union of two intervals allow us to correctly label a point set of the form $-+-+-$. All labelings for 4 points can be easily shown to be consistent with this label format. For any arrangement of 5 points, however, there exists the labeling $+-+-+$ which cannot be accomplished by the union of two intervals.

3. Origin centred circle binary classifiers: $f : \mathbb{R}^2 -> \{0, 1\}$, $b > 0$, where example is labeled positive if it lies inside the circle, and negative otherwise.

$$H = \{f(x) = I_{wx^T x \leq b}\}$$

Solution VCdim = 1. For one-point, we can classify the each of the two possible labelling, so VCdim is at least 1. Consider any two points on the plane p_1 and p_2. Assume that $||p_1||_2 < ||p_2||_2$ and p_1 has label -1 and p_2 – 1. Then we can not find a circle centered at the origin and only include p_2.

4. Origin centred circle binary classifiers given in 3 and the functions that flip the outputs of the functions in 3.

Solution VCdim = 2. See Figure 1 to prove it is at least 2. Consider any three points on the plane p_1 and p_2 and p_3. Assume that $||p_1||_2 < ||p_2||_2 < ||p_3||_2$ and p_1 has label -1, p_2 – 1, p_3 – -1 (or 1, -1, 1 accordingly). Then we can not find a circle centered at the origin that ensures such labeling.

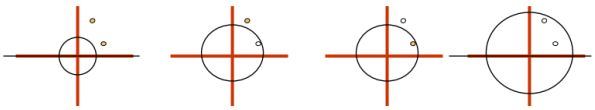

Figure 1: VCdim of circles

5. A set of 3-node decision trees in one dimension $\mathbb{R}$.

Solution VCdim = 4. Each node defines a newsplitting threshold on the data. 3 thresholds define at most 4 regions on the axis, so the CV dimension is at least 4. The same argument as in the closed intervals case is used to prove that VC dimension is less than 5.

6. A set of axis-parallel squares in $\mathbb{R}^2$. Point is labeled positive if it lies inside the square, and negative otherwise.

$$H = \{f(x) = I_{max(|x_1|,|x_2|)=c}\}$$

Solution VCdim = 3. It is not hard to see that the set of 3 points with coordinates (1, 0), (0, 1), and (-1, 0) can shattered by axis-aligned squares: e.g., to label positively two of these points, use a square defined by the axes and with those to points as corners. Thus, the VC-dimension is at least 3. No set of 4 points can be fully shattered. To see this, let PT be the highest point, PB the lowest, PL the leftmost, and PR the rightmost, assuming for now that these can be defined in a unique way (no tie) – the cases where there are ties can be treated in a simpler fashion. Assume without loss of generality that the difference dBT of y-coordinates between PT and PB is greater than the difference dLR of x-coordinates between PL and PR. Then, PT and PB cannot be labeled positively while PL and PR are labeled negatively. Thus, the VC-dimension of axis-aligned squares in the plane is 3.

7. A system of all convex polygons in $\mathbb{R}^2$. Point is labeled positive if it lies inside or on the edge of the convex polygon, and negative otherwise.

$$H = \{f(x) = I_{x \in C} \ | C \text{ convex in } \mathbb{R}^2\}$$

Solution VCdim = ∞. For any n points let consider an arrangement where all points are on the unit circle. Then any positive subset of the points can be vertexes of a convex polygon C_n. This polygon also does not contain any negative point as they are on a circle. Since we can do this arrangements for any n, the cardinality of the shattered subsets is unbounded.

8. Finite hypothesis space H. Prove that VC dimension of a finite hypothesis space H is upper bounded by $\log_2 |H|$.

   Solution Assume that VC dimension of a finite hypothesis space H is d. Then by definition of VC dimension there is a set S of a d distinct points such that S can be shattered. For a set of size d there exists 2^d distinct labeling concepts, therefore H should contain at least 2^d different hypotheses (each concept gets a different hypothesis). Therefore $2^d \leq |H|$ and we get that $d \leq \log_2 |H|$

3 Assorted Questions – SVM, Kernels, Convexity, Logistic regression

1. A ML lover student trains an SVM on a particular set of training data. If student then adds a new training point to the dataset and retrains the SVM, the number of support vectors may. **Answer**. All of them

2. We would expect the support vectors to remain the same in general as we move from a linear kernel to higher order polynomial kernels.

   Answer. False.

3. VC Dimension depends on the dataset we use for shattering.

   Answer. False.

4. The maximum margin decision boundaries that support vector machines construct have the lowest generalization error among all linear classifiers.

   Answer. False.

5. Following constrained optimization problem is equivalent to optimization problem solved by SVM

$$\max_{\lambda,\lambda_0,\gamma} \gamma \text{ s.t. } y_i \frac{\lambda^T x_i + \lambda_0}{||\lambda||} \geq \gamma, \quad i = 1, ..., n$$

   Answer. True.

6. If the VC Dimension of a set of classification hypotheses is ∞, then the set of classifiers can achieve 100% training accuracy on any dataset.

   Answer. True. If the VC Dimension of a set of classification hypotheses is infinite, it can shatter some dataset D at any given dataset size. Finding a mapping from a given dataset D0 to D allows 100% training accuracy on any dataset

7. Since the true risk is bounded by the empirical risk, it is a good idea to minimize the training error as much as possible. **Answer**. False.

8. VC Dimensions of the sets of classification hypotheses induced by logistic regression and linear SVM (learnt on the same set of features) are different. **Answer**. False. They both induce linear separators. Since the set of classification hypotheses is the same, the VC Dim is the same.

9. The Gram matrix $G = \mathbf{1}\mathbf{1}^\top$ where $\mathbf{1}$ is the all-1 vector is positive semi-definite.

   Answer. True.

10. We can use the kernel trick to Logistic regressions.

    Answer. True. It is called kernel logistic regression.

11. Consider a SVM with the Gaussian RBF kernel: $k(xi, xj) = exp(-\frac{-||x_i - x_j||^2}{2\sigma})$. Suppose we have three points, z1, z2, and x. z1 is geometrically very close to x, and z2 is geometrically far away from x. What is the value of $k(z1, x)$?

    Answer. RBF kernel generates a "bump" around the center x. For points z1 close to the center of the bump, K(z1, x) will be close to 1.

12. Consider a SVM with the Gaussian RBF kernel: $k(xi, xj) = exp(-\frac{-||x_i - x_j||^2}{2\sigma})$. Suppose we have three points, z1, z2, and x. z1 is geometrically very close to x, and z2 is geometrically far away from x. What is the value of $k(z2, x)$?

    Answer. For points away from the center of the bump K(z2, x) will be close to 0.

13. Assume that $k_1(x, x')$ and $k_2(x, x')$ are valid kernel, them kernel $k_1(x, x') - k_2(x, x')$ is also valid.

    Answer. False

14. Assume that $k(x, x')$ is a valid kernels, then it is true that $k(x, y) \leq k(x, x)k(y, y)$ **Answer**. True

$$k(x, y)^2 =< \phi(x), \phi(y) >^2 = ||\phi(x)||^2||\phi(y)||^2 cos\theta^2_{\phi(x),\phi(y)} \leq ||\phi(x)||^2||\phi(y)||^2 = k(x, x)k(y, y)$$

15. The Cobb-Douglas production function is widely used in economics to represent the relationship between inputs and outputs of a firm. It takes the form $Y = AL^\alpha K^\beta$ where Y represents output, L labor, and K capital. The parameters α and β are constants that determine how production is scaled. The Cobb-Douglas function can also be applied to utility maximization and takes the general form $\prod_{i=1}^{N} x_i^{\alpha_i}$. Consider the following utility maximization problem:

$$\max_x u(x) = x_1^\alpha x_1^{1-\alpha} \text{ s.t.}$$

$$p_1 x_1 + p_2 x_2 \leq w \quad x_1, x_2 \geq 0$$

    Can we turn this problem into a minimization problem with convex objective?

    Answer. True. $u'(x) = -\alpha \log x_1 - (1 - \alpha) \log x_2$