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.

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.

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.

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:

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

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.

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

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.

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

4. Origin centred circle binary classifiers given in 3 and the functions that flip the outputs of the functions in 3.
5. A set of 3-node decision trees in one dimension $\mathbb{R}$.
6. A set of axis-parallel squares in $\mathbb{R}^2$. Point is labeled positive if it lies inside the square, and negative otherwise.

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

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.

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

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

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.

Increase	Decrease
Stay the same	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.

True	False

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

True	False

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

True	False

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

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

True	False

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.

True	False

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.

True	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.

True	False

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

True	False

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

True	False

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 $k(z1, x)$?

$k(z_1, x)$ will be close to 1	$k(z_1, x)$ will be close to 0
$k(z_1, x)$ will be close to c_1, $c_1 >> 1, c_1 \in \mathbb{R}$	$k(z_1, x)$ will be close to c_1, $c_1 << 0, c_1 \in \mathbb{R}$

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 $k(z2, x)$?

$k(z_2, x)$ will be close to 1	$k(z_2, x)$ will be close to 0
$k(z_2, x)$ will be close to c_2, $c_2 >> 1, c_1 \in \mathbb{R}$	$k(z_2, x)$ will be close to c_2, $c_2 << 0, c_2 \in \mathbb{R}$

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.

True	False

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

True	False

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?

True	False