Avisek Gupta (avisek003@gmail.com)

Machine Learning

18 – VC Dimensions, Support Vector Machines (1)

November 17, 2022

PAC Learning

Theorem (Valiant, 1984): If the hypothesis space H is finite, and D is a sequence of $n \geq 1$ independent random examples of some target concept c, then for any $0 \leq \varepsilon \leq 1$, the probability that $VS_{H,D}$ contains a hypothesis with error greater than ε is less than $|H|e^{-\varepsilon n}$, i.e.,

$$Pr[Err > \varepsilon] < |H|e^{-\varepsilon n}$$

Let us want this probability to be at most δ, i.e.,

$$|H|e^{-\varepsilon n} \leq \delta$$

Then,

$$n \geq \frac{1}{\varepsilon}(\ln |H| + \ln(1/\delta))$$

1. With linear increase in data, the bound becomes exponentially better.
2. $|H|$ can be large, requiring more data (If $|H|$ is infinity, the bound does not help).

PAC Learning

Theorem (Valiant, 1984): If the hypothesis space H is finite, and D is a sequence of $n \geq 1$ independent random examples of some target concept c, then for any $0 \leq \varepsilon \leq 1$, the probability that $VS_{H,D}$ contains a hypothesis with error greater than ε is less than $|H|e^{-\varepsilon n}$, i.e.,

$$Pr[Err > \varepsilon] < |H|e^{-\varepsilon n}$$

Let us want this probability to be at most δ, i.e.,

$$|H|e^{-\varepsilon n} \leq \delta$$

Then,

$$n \geq \frac{1}{\varepsilon}(\ln |H| + \ln(1/\delta))$$

1. With linear increase in data, the bound becomes exponentially better.
2. $|H|$ can be large, requiring more data (If $|H|$ is infinity, the bound does not help). [What can we do if $|H|$ is infinity?]

VC Dimensions

Dichotomy: A dichotomy of a set S is a partition of S into two disjoint subsets.

VC Dimensions

Dichotomy: A dichotomy of a set S is a partition of S into two disjoint subsets.

Shattering: A set of instances S is said to be *shattered* by a hypothesis space H iff for every dichotomy of S, there exists some hypothesis in H consistent with this dichotomy.

VC Dimensions

Dichotomy: A dichotomy of a set S is a partition of S into two disjoint subsets.

Shattering: A set of instances S is said to be *shattered* by a hypothesis space H iff for every dichotomy of S, there exists some hypothesis in H consistent with this dichotomy.

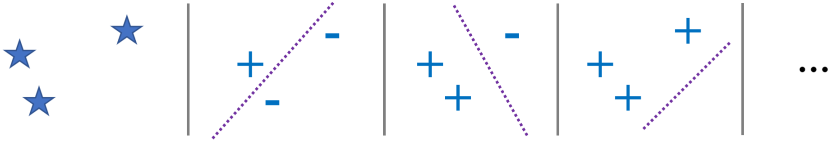


Figure: Shattering a set S with $|S| = 3$ by a set of hypotheses of staight lines.

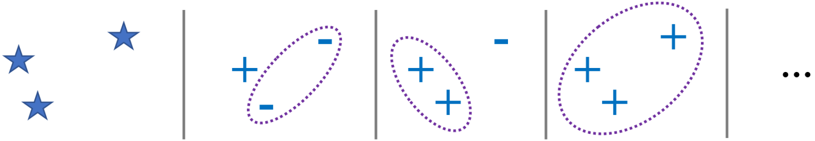


Figure: Shattering a set S with $|S| = 3$ by a set of hypotheses of ellipses.

VC Dimensions

Dichotomy: A dichotomy of a set S is a partition of S into two disjoint subsets.

Shattering: A set of instances S is said to be *shattered* by a hypothesis space H iff for every dichotomy of S, there exists some hypothesis in H consistent with this dichotomy.

Vapnik-Chervonenkis Dimension (1971): $VC(H)$ of a hypothesis space H defined over instance space X is the size of the largest finite subset of X shattered by H.

If arbitrary large finite sets of X can be shattered by H, then $VC(H) \equiv \infty$.

VC Dimensions

Dichotomy: A dichotomy of a set S is a partition of S into two disjoint subsets.

Shattering: A set of instances S is said to be *shattered* by a hypothesis space H iff for every dichotomy of S, there exists some hypothesis in H consistent with this dichotomy.

Vapnik-Chervonenkis Dimension (1971): $VC(H)$ of a hypothesis space H defined over instance space X is the size of the largest finite subset of X shattered by H.

If arbitrary large finite sets of X can be shattered by H, then $VC(H) \equiv \infty$.

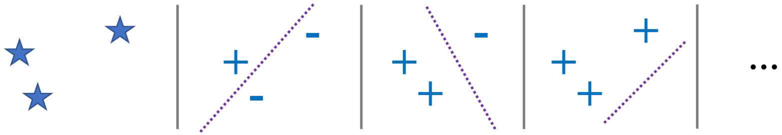


Figure: 1. Shattering $\mathbb{R}^2$ with $n = 3$ by a set of hypotheses of staight lines.

VC Dimensions

Dichotomy: A dichotomy of a set S is a partition of S into two disjoint subsets.

Shattering: A set of instances S is said to be *shattered* by a hypothesis space H iff for every dichotomy of S, there exists some hypothesis in H consistent with this dichotomy.

Vapnik-Chervonenkis Dimension (1971): $VC(H)$ of a hypothesis space H defined over instance space X is the size of the largest finite subset of X shattered by H.

If arbitrary large finite sets of X can be shattered by H, then $VC(H) \equiv \infty$.

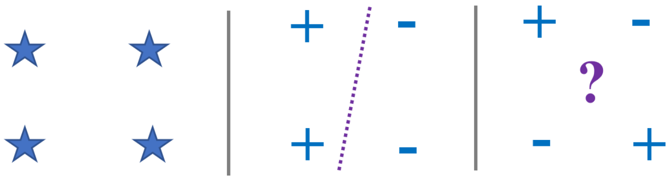


Figure: 2. $\mathbb{R}^2$ with $n = 4$ cannot be shattered by a set of hypotheses of staight lines.

VC Dimensions

Dichotomy: A dichotomy of a set S is a partition of S into two disjoint subsets.

Shattering: A set of instances S is said to be *shattered* by a hypothesis space H iff for every dichotomy of S, there exists some hypothesis in H consistent with this dichotomy.

Vapnik-Chervonenkis Dimension (1971): $VC(H)$ of a hypothesis space H defined over instance space X is the size of the largest finite subset of X shattered by H.

If arbitrary large finite sets of X can be shattered by H, then $VC(H) \equiv \infty$.

For $\mathbb{R}^2$ and a hypothesis set H of straight lines:

- $n = 1$ is shattered by H.
- $n = 2$ is shattered by H.
- $n = 3$ is shattered by H.
- $n = 4$ is **not** shattered by H.

Hence $VC(H) = 3$.

VC Dimensions

Dichotomy: A dichotomy of a set S is a partition of S into two disjoint subsets.

Shattering: A set of instances S is said to be *shattered* by a hypothesis space H iff for every dichotomy of S, there exists some hypothesis in H consistent with this dichotomy.

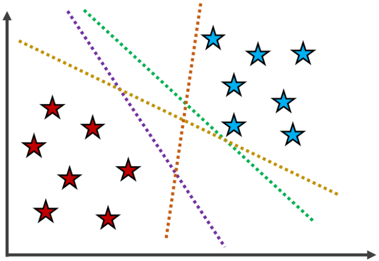
Vapnik-Chervonenkis Dimension (1971): $VC(H)$ of a hypothesis space H defined over instance space X is the size of the largest finite subset of X shattered by H.

Sample Complexity from VC Dimension (2000): The number of randomly drawn examples that suffice to guarantee error of at most ε with probability at least $(1 - \delta)$ is:

$$n \geq \frac{1}{\varepsilon}(8\,VC(H)\,\ln(13/\varepsilon) + 4\,\ln(2\delta))$$

Support Vector Machines

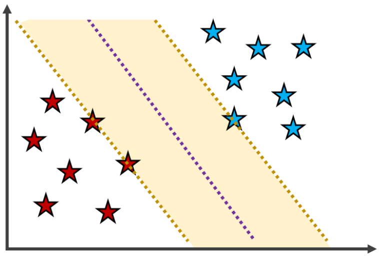
Classification Hyperplanes



- Given sample instances from two linearly separable classes, there is an infinite number of hyperplanes that can correctly classify the samples.
- Can we create a definition of an *ideal* hyperplane?

Maximum-Margin Hyperplanes

An *ideal* hyperplane: A hyperplane that has the **maximum margin** between the two classes.

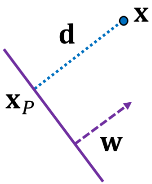


How can we define the margin of a hyperplane?

Maximum-Margin Hyperplanes

Notations: For a binary classification problem, we have samples $(\mathbf{x}^{(1)}, y^{(1)}), ..., (\mathbf{x}^{(n)}, y^{(n)}), \mathbf{x}^{(i)} \in \mathbb{R}^d, y^{(i)} \in \{1, -1\}$.

We wish to find a hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ that has maximum margin, and correctly classifies the data.



We can define the distance vector $\mathbf{d}$ of any instance $\mathbf{x}$ to a hyperplane.

The normal to the hyperplane is $\mathbf{w}$, hence $\mathbf{d} = \alpha\mathbf{w}$.

Let the projection of $\mathbf{x}$ on to the hyperplane be $\mathbf{x}_P$, therefore $\mathbf{x}_P = \mathbf{x} - \mathbf{d}$. Then we can solve for α,

$$\mathbf{w}^T\mathbf{x}_P + b = 0$$

$$\implies \mathbf{w}^T(\mathbf{x} - \mathbf{d}) + b = 0$$

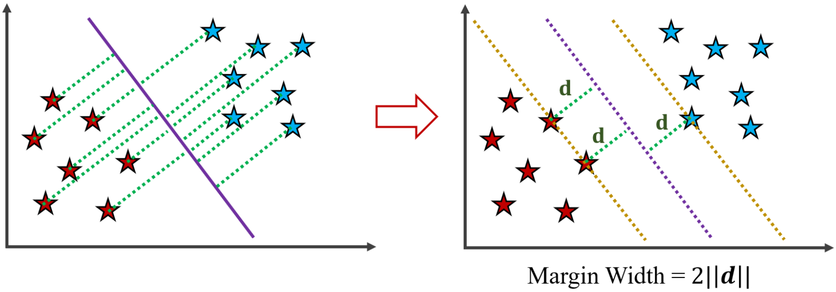
$$\implies \mathbf{w}^T(\mathbf{x} - \alpha\mathbf{w}) + b = 0$$

$$\implies \alpha = \frac{\mathbf{w}^T\mathbf{x} + b}{\mathbf{w}^T\mathbf{w}}$$

Maximum-Margin Hyperplanes

For any $\mathbf{x}$, the distance vector to a hyperplane vector $\mathbf{d} = \alpha \mathbf{w}, \alpha = \dfrac{\mathbf{w}^T \mathbf{x} + b}{\mathbf{w}^T \mathbf{w}}$

Since we are interested in finding the *maximum* margin hyperplane, we can try to find the instances that are *nearest* to the hyperplane. The norm of their distance vectors will provide a measure of the margin width.



Maximum-Margin Hyperplanes

For any $\mathbf{x}$, the distance vector to a hyperplane vector $\mathbf{d} = \alpha \mathbf{w}, \ \alpha = \dfrac{\mathbf{w}^T \mathbf{x} + b}{\mathbf{w}^T \mathbf{w}}$

$$\begin{aligned} ||\mathbf{d}||_2 &= \sqrt{\mathbf{d}^T \mathbf{d}} = \sqrt{\alpha^2 \mathbf{w}^T \mathbf{w}} \\ &= \frac{|\mathbf{w}^T \mathbf{x} + b|}{\sqrt{\mathbf{w}^T \mathbf{w}}} = \frac{|\mathbf{w}^T \mathbf{x} + b|}{||\mathbf{w}||_2} \end{aligned}$$

The margin γ is then defined as,

$$\gamma(\mathbf{w}, b) = \min_{\mathbf{x}} \ \frac{2|\mathbf{w}^T \mathbf{x} + b|}{||\mathbf{w}||_2}$$

Note that by this definition, the margin is scale invariant:

$$\gamma(\beta \mathbf{w}, \beta b) = \gamma(\mathbf{w}, b), \ \forall \beta \neq 0$$

Maximum-Margin Hyperplanes

Finding the maximum margin hyperplane can be posed as an optimization problem,

$$\max_{\mathbf{w},b} \; \gamma(\mathbf{w}, b) \;\; \text{s.t.} \;\; y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 0 \;\; \forall i$$

Note that since $y^{(i)} \in \{1, -1\}$, an accurate classifier will have:

$$y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 0 \;\; \forall i.$$

Maximum-Margin Hyperplanes

Finding the maximum margin hyperplane can be posed as an optimization problem,

$$\max_{\mathbf{w},b} \ \gamma(\mathbf{w}, b) \ \text{ s.t. } \ y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 0 \ \ \forall i$$

Equivalently, the objective is,

$$\max_{\mathbf{w},b} \left\{ \frac{2}{||\mathbf{w}||_2} \ \min_{\mathbf{x}^{(i)}} \ |\mathbf{w}^T\mathbf{x}^{(i)} + b| \right\} \ \text{ s.t. } \ y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 0 \ \ \forall i$$

We impose the constraint that $\min_{\mathbf{x}} |\mathbf{w}^T\mathbf{x} + b| = 1$ to prevent an arbitrary large solution. Then the optimization objective is,

$$\max_{\mathbf{w},b} \frac{2}{||\mathbf{w}||_2}.1 = \min_{\mathbf{w},b} \ ||\mathbf{w}||_2 = \min_{\mathbf{w},b} \ \mathbf{w}^T\mathbf{w}$$

$$\text{s.t., } y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 0 \ \forall i,$$

$$\min_{i} \ |\mathbf{w}^T\mathbf{x}_i + b| = 1.$$

Support Vector Machines

The optimization objective:

$$\max_{\mathbf{w},b} \frac{2}{||\mathbf{w}||_2}.1 = \min_{\mathbf{w},b} \; ||\mathbf{w}||_2 = \min_{\mathbf{w},b} \; \mathbf{w}^T\mathbf{w}$$

$$\text{s.t., } y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 0 \; \forall i,$$

$$\min_{i} \; |\mathbf{w}^T\mathbf{x}^{(i)} + b| = 1.$$

Combining the constraints, we get an objective function for the classifier called Support Vector Machines:

$$\min_{\mathbf{w},b} \; \mathbf{w}^T\mathbf{w}$$

$$\text{s.t., } y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 \; \forall i.$$

Support Vector Machines

The optimization objective:

$$\max_{\mathbf{w},b} \frac{2}{||\mathbf{w}||_2}.1 = \min_{\mathbf{w},b} \; ||\mathbf{w}||_2 = \min_{\mathbf{w},b} \; \mathbf{w}^T\mathbf{w}$$

$$\text{s.t., } y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 0 \; \forall i,$$

$$\min_{i} \; |\mathbf{w}^T\mathbf{x}^{(i)} + b| = 1.$$

Combining the constraints, we get an objective function for the classifier called Support Vector Machines:

$$\min_{\mathbf{w},b} \; \mathbf{w}^T\mathbf{w}$$

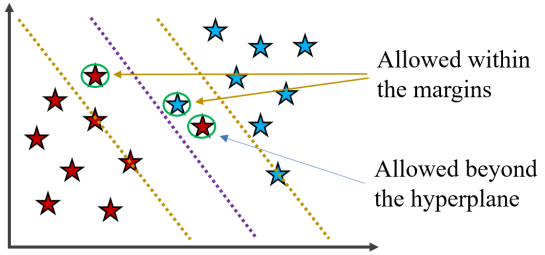
$$\text{s.t., } y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 \; \forall i.$$

Quadratic Optimization Problem (QOP): An optimization problem with a quadratic objective and with linear equality or linear inequality constraints. Quadratic solvers can solve QOPs (but they are not very efficient).

Support Vector Machines with soft constraints

Our initial assumptions: Classes are linearly separable.

What if the classes are not linearly separable? We can find a maximum margin classifier that allows some misclassification.



Slack variables $\xi^{(i)}$ are imposed to allow instances to cross the margin:

$$y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - \xi^{(i)} \ \forall i$$

Support Vector Machines with soft constraints

Soft SVM Objective:

$$\min_{\mathbf{w},b} \; \mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi^{(i)}$$

$$\text{s.t., } y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - \xi^{(i)} \;\; \forall i,$$

$$\xi^{(i)} \geq 0 \; \forall i.$$

C controls how strict SVM is too get all points on the correct side of the hyperplane. For larger C, SVM will try to be very accurate. For smaller C, SVM will allow more points to be on the incorrect side of the hyperplane.

Support Vector Machines with soft constraints

Soft SVM Objective:

$$\min_{\mathbf{w},b} \ \mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n} \xi^{(i)}$$

$$\text{s.t., } y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - \xi^{(i)} \ \forall i,$$

$$\xi^{(i)} \geq 0 \ \forall i.$$

For points that satisfy the constraint $y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1$, setting $\xi^{(i)} = 0$ minimizes the objective.

Only for $y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) < 1$, do we need to set a proper $\xi^{(i)} > 0$ so that the constraint $y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - \xi^{(i)}$ is satisfied. The minimum value of $\xi^{(i)}$ that satisfies the constraint is then,

$$\xi^{(i)} = \begin{cases} 1 - y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) & \text{, if } y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) < 1 \\ 0 & \text{, if } y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 \end{cases}$$

Support Vector Machines with soft constraints

Soft SVM Objective:

$$\min_{\mathbf{w},b} \; \mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi^{(i)}$$
$$\text{s.t., } y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - \xi^{(i)} \;\; \forall i,$$
$$\xi^{(i)} \geq 0 \; \forall i.$$

A solution for $\xi^{(i)}$,

$$\xi^{(i)} = \begin{cases} 1 - y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) & \text{, if } y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) < 1 \\ 0 & \text{, if } y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 \end{cases}$$

Equivalently,

$$\xi^{(i)} = \max(1 - y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b), 0)$$

Support Vector Machines with soft constraints

Soft SVM Objective:

$$\min_{\mathbf{w},b} \ \mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n} \xi^{(i)}$$

$$\text{s.t., } y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - \xi^{(i)} \ \forall i,$$

$$\xi^{(i)} \geq 0 \ \forall i.$$

A solution for $\xi^{(i)}$,

$$\xi^{(i)} = \max(1 - y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b), 0)$$

A combined objective for SVM with Soft Constraints:

$$\min_{\mathbf{w},b} \ \mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n} \max\left[1 - y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b), 0\right]$$

Support Vector Machines with soft constraints

Soft SVM Objective:

$$\min_{\mathbf{w},b} \underbrace{\mathbf{w}^T\mathbf{w}}_{\ell_2 \text{ regularizer}} + C \underbrace{\sum_{i=1}^{n} \max\left[1 - y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b), 0\right]}_{\text{Hinge Loss}}$$

This is similar to a Logistic Regression objective function with an ℓ_2 penalty.

References

(Classification Validation) [1] Mohammed J. Z. and Wagner M. (2020). Data Mining and Machine Learning: Fundamental Concepts and Algorithms, Second Edition. Cambridge University Press.

(No Free Lunch Theorem) [2] Duda R. O., Hart P.E., Stork D. G. (2000). Chapter 9 - 9.2. No Free Lunch Theorem. Pattern Classification, 2nd Edition.

(PAC Learning) [3] Valiant L. (1984). "A theory of the learnable". Communications of the ACM. 27 (11): 1134-1142.

(VC Dimensions) [4] Vapnik V. N. and Chervonenkis A. Y. (1971). "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities". Theory of Probability & Its Applications. 16 (2): 264.

(VC Dimensions) [5] Vapnik V. N. (2000). The Nature of Statistical Learning Theory. Springer.

(Agnostic Learning) [6] Ali Ghodsi, Lec 19: PAC Learning. URL: `https://www.youtube.com/watch?v=qOMOYMOWCzU`.

(SVM) [7] Kilian Q. Weinberger, CS4780/CS5780 Lecture 9. URL: `https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote09.html`

(SVM) [8] Andrew Ng, CS229 Support Vector Machines. URL: `https://see.stanford.edu/materials/aimlcs229/cs229-notes3.pdf`