Generic support vector machines and Radon's theorem

Brittany Carr
Advisor: Dr. Henry Adams
Committee: Dr. Patrick Shipman and Dr. Anders Fremstad

May 1, 2019

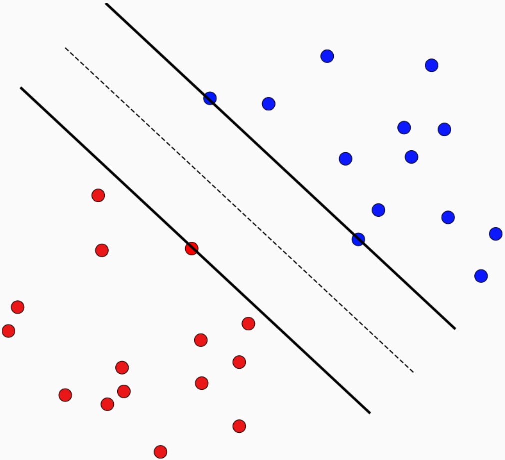
Colorado State University

Outline

- Overall purpose
- Support vector machines
- Radon's theorem
- Classical algebraic geometry
- Support vector configurations
- Acknowledgements

Support Vector Machines

Goal: Understand the possible geometric configurations of support vectors



Support Vector Machines

The classifier is given by $f(x) = w^T x + b$ where

- x: The data points
- b: Shift of the hyperplane away from the origin
- w: The normal vector defining the hyperplane
- $y_x \in \{-1, 1\}$: Labels for our data (used later)

Support Vector Machines

The classifier is given by $f(x) = w^T x + b$ where

- x: The data points
- b: Shift of the hyperplane away from the origin
- w: The normal vector defining the hyperplane
- $y_x \in \{-1, 1\}$: Labels for our data (used later)

We have three possibilities for $f(x)$:

- $f(x) = 0$: Defines the separating hyperplane
- $|f(x)| = 1$: Defines the support vectors
- $|f(x)| > 1$: Other data points farther away from the separating hyperplane

Note, $sign(f(x))$ determines which class the data is in

Requirements for hard margin SVM:

- No points can lie inside the margin
- Linearly separable data; no misclassification
- Data points in $\mathbb{R}^n$

Support Vector Machines

Mathematically, SVM is an optimization problem. We want to **maximize** the margin of the separating hyperplane where the margin is $\frac{2}{||w||}$.

Mathematically, SVM is an optimization problem. We want to **maximize** the margin of the separating hyperplane where the margin is $\frac{2}{||w||}$.

Thus, we **minimize** $\frac{1}{2}||w||^2$

$$\underset{\mathbf{w},b}{\arg\min} \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } y_i \left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 \text{ for all } i.$$

The SVM dual and the KKT conditions

Theorem (Karush-Kuhn-Tucker)

Consider an optimization problem in $\mathbb{R}^n$ *of the form*

$$\mathbf{min}(f(x)) \textit{ subject to } g_i(x) \leq 0 \textit{ for all } i = 1, \ldots, m,$$

where $f(x)$ *is a differentiable function of input variables x, and the* $g_i(x)$ *are affine degree one polynomials. Suppose z is a local minimum of f. Then, there exist constants* $\alpha_1, \alpha_2, \ldots, \alpha_m \in \mathbb{R}$ *such that*

(1) $-\nabla f(z) = \sum_{i=1}^{m} \alpha_i \nabla g_i(z)$ ***The Lagrangian is 0***

(2) $g_i(z) \leq 0$ *for all i,* ***Gives us that the original constraints are satisfied***

(3) $\alpha_i \geq 0$ *for all i, and* ***Gives us that the dual constraints are satisfied***

(4) $\alpha_i g_i(z) = 0$ *for all i.* ***Support vectors have margin exactly 1***

After translating into the dual we have

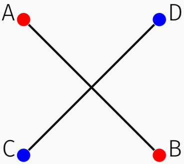
$$
\begin{aligned}
L(\mathbf{w}, b, \alpha) &= \frac{1}{2}||\mathbf{w}||^2 + \sum_{j=1}^{m} y_j \alpha_j (\langle \mathbf{w}, \mathbf{x}_j \rangle + b) \\
&= \frac{1}{2}||\mathbf{w}||^2 + \sum_{j=1}^{m} \alpha_j - \sum_{j=1}^{m} y_j \alpha_j (\langle \mathbf{w}, \mathbf{x}_j \rangle) - \sum_{j=1}^{m} \alpha_j y_j b.
\end{aligned}
$$

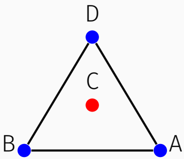
Utilizing the KKT conditions, this problem simplifies into

$$
L(\alpha) = \sum_{j=1}^{m} \alpha_j - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle,
$$

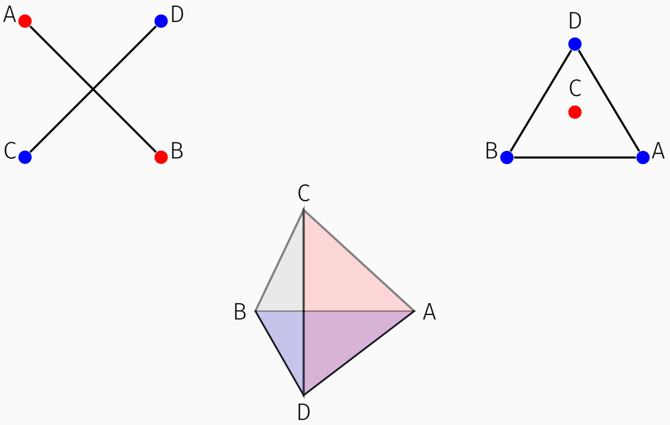
which is an unbounded maximization problem.

Radon's Theorem: If T is a set of k points in Euclidean n-dimensional space $\mathbb{R}^n$ with $k \geq n+2$, then there exist disjoint sets T_1 and T_2 with $T = T_1 \cup T_2$ and $\mathrm{conv}(T_1) \cap \mathrm{conv}(T_2) \neq \varnothing$.



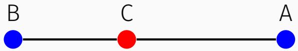


Radon's Theorem: If T is a set of k points in Euclidean n-dimensional space $\mathbb{R}^n$ with $k \geq n+2$, then there exist disjoint sets T_1 and T_2 with $T = T_1 \cup T_2$ and $\mathrm{conv}(T_1) \cap \mathrm{conv}(T_2) \neq \varnothing$.



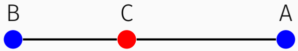
Radon's Theorem

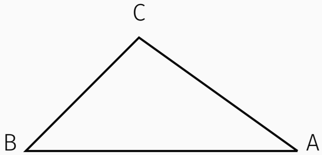
Radon's Theorem: If T is a set of k points in Euclidean n-dimensional space $\mathbb{R}^n$ with $k \geq n+2$, then there exist disjoint sets T_1 and T_2 with $T = T_1 \cup T_2$ and $\mathrm{conv}(T_1) \cap \mathrm{conv}(T_2) \neq \varnothing$.



Radon's Theorem

Radon's Theorem: If T is a set of k points in Euclidean n-dimensional space $\mathbb{R}^n$ with $k \geq n + 2$, then there exist disjoint sets T_1 and T_2 with $T = T_1 \cup T_2$ and $\mathrm{conv}(T_1) \cap \mathrm{conv}(T_2) \neq \varnothing$.

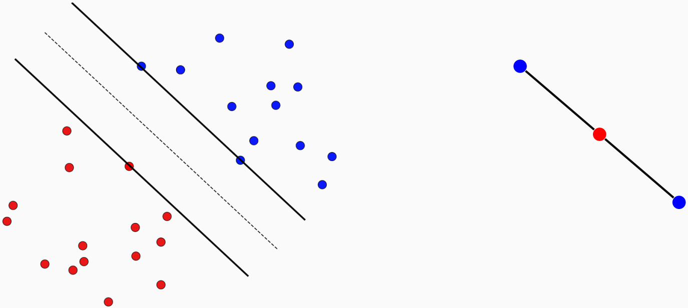




Lemma

If $X \subset \mathbb{R}^n$ is a set of linearly separable labeled points, then the projections of the convex hulls of the positive and negative support vectors onto the separating hyperplane intersect.

Radon's theorem and SVM configurations



For points in "general position", we want to show there is only one Radon point.

Algebraic varieties

Definition

An *affine variety* is the set of common zeros of a finite family of polynomials. Given a set $S \subseteq \mathbb{A}[x_1, x_2, \cdots x_n]$ of polynomials in some affine space $\mathbb{A}^n$, the affine variety defined by S is the set

$$\mathcal{V}(S) := \{a \in \mathbb{A}^n \mid f(a) = 0 \text{ for all } f \in S\}.$$

Definition

An *affine variety* is the set of common zeros of a finite family of polynomials. Given a set $S \subseteq \mathbb{A}[x_1, x_2, \cdots x_n]$ of polynomials in some affine space $\mathbb{A}^n$, the affine variety defined by S is the set

$$\mathcal{V}(S) := \{a \in \mathbb{A}^n \mid f(a) = 0 \text{ for all } f \in S\}.$$

Theorem

The intersection of any collection of affine varieties is an affine variety. The union of any finite collection of affine varieties is an affine variety.

Definition

The *determinant* of an $n \times n$ matrix A is

$$det(A) = \sum_{\sigma \in S_n} \left(sgn(\sigma) \prod a_{i,\sigma_i} \right),$$

where σ is an element in the symmetric group on n elements, and a_{i,σ_i} represents the ith row and σ_ith column entry of A.

Example

Let $A = [a_{i,j}]$ where $1 \leq i \leq 3$ and $1 \leq j \leq 3$. Thus,

$$\begin{aligned} det(A) &= \sum_{\sigma \in S_3} \left(sgn(\sigma) \prod a_{i,\sigma_i} \right) \\ &= sgn(e) \prod a_{i,(e)_i} + sgn(123) \prod a_{i,(123)_i} + sgn(132) \prod a_{i,(132)_i} \\ &\quad + sgn(13) \prod a_{i,(13)_i} + sgn(12) \prod a_{i,(12)_i} + sgn(23) \prod a_{i,(23)_i} \\ &= \prod a_{i,(e)_i} + \prod a_{i,(123)_i} + \prod a_{i,(132)_i} - \prod a_{i,(13)_i} - \prod a_{i,(12)_i} - \prod a_{i,(23)_i} \\ &= a_{1,1}a_{2,2}a_{3,3} + a_{1,2}a_{2,3}a_{3,1} + a_{1,3}a_{2,1}a_{3,2} \\ &\quad - a_{1,3}a_{2,2}a_{3,1} - a_{1,2}a_{2,1}a_{3,3} - a_{1,1}a_{2,3}a_{3,2}. \end{aligned}$$

If $M(y)$ is an $m \times n$ matrix with $m \geq n$ with linear (or even polynomial) functions in the entries of y, then the determinants of all $n \times n$ minors gives a collection of $\binom{m}{n}$ polynomial functions.

$$\begin{bmatrix} f_{1,1}(y) & f_{1,2}(y) & \cdots & f_{1,n}(y) \\ f_{2,1}(y) & f_{2,2}(y) & \cdots & f_{2,n}(y) \\ \vdots & \vdots & \ddots & \vdots \\ f_{n,1}(y) & f_{n,2}(y) & \cdots & f_{n,n}(y) \\ \vdots & \vdots & \ddots & \vdots \\ f_{m,1}(y) & f_{m,2}(y) & \cdots & f_{m,n}(y) \end{bmatrix}$$

Further if our matrix is rank deficient, then the determinants of all $n \times n$ minors are zero. Hence we have a collection of polynomials set equal to zero and we can use them to define an algebraic variety.

Definition

Let $\mathcal{V}_{\mathrm{rd}}(m, n) \subseteq \mathbb{R}^{mn}$ for $m \geq n$ be the algebraic variety generated by the set of polynomials $\{A_i\}_{i \in I}$, where I is the set containing all $\binom{m}{n}$ choices of n rows, and A_i is the minor of the submatrix consisting of those rows.

Lemma

Let $M(y)$ be an $m \times n$ matrix with $m \geq n$, depending on $y \in \mathbb{R}^k$. Suppose the entries of $M(y)$ are linear (or even polynomial) functions in the entries of y. Then $\mathcal{V}_{M(y)} := \{y \in \mathbb{R}^k \mid M(y) \text{ is rank deficient}\}$ is an algebraic variety.

Radon’s theorem and SVM configurations

Using the varieties we have defined above, we can say that a set of points in general position is open and dense in the Euclidean topology, ”aka generic”.

Radon's theorem and SVM configurations

Using the varieties we have defined above, we can say that a set of points in general position is open and dense in the Euclidean topology, ”aka generic”.

We can also say that a set of points in strong general position is open and dense in the Euclidean topology, ”aka generic”.

Radon's theorem and SVM configurations

Using the varieties we have defined above, we can say that a set of points in general position is open and dense in the Euclidean topology, ”aka generic”.

We can also say that a set of points in strong general position is open and dense in the Euclidean topology, ”aka generic”.

Thus we can perturb all points by some ε such that they remain in (strong) general position.

Definition

A set of points $X \subseteq \mathbb{R}^n$ is in *strong general position* if

(i) for $k < n$, no $k+2$ subset of X lies in a k-flat

(ii) for any $k+1$ points in X (determining a k-flat), the orthogonal projection of any other point in X to that k-flat does not hit the affine span of k of those points

(iii) for $k + l \leq n$, no disjoint k-flats and l-flats contain parallel vectors.

Radon's theorem and SVM configurations

Lemma

If $X \subset \mathbb{R}^n$ *is a set of linearly separable labeled points in strong general position, then the projections of the convex hulls of the positive and negative support vectors onto the separating hyperplane intersect at a single Radon point.*

Radon's theorem and SVM configurations

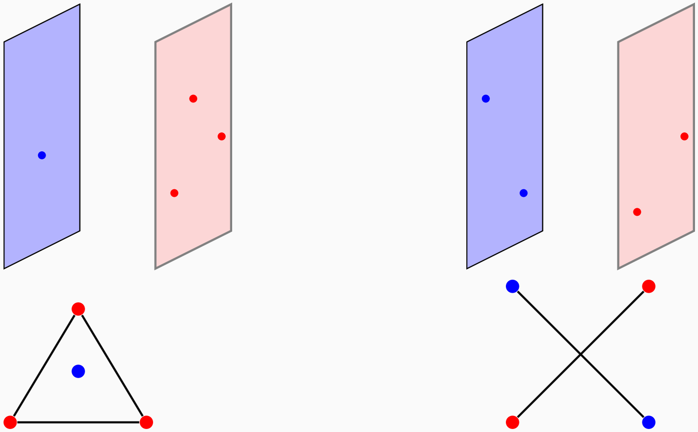
Lemma

If $X \subset \mathbb{R}^n$ is a set of linearly separable labeled points in strong general position, then the projections of the convex hulls of the positive and negative support vectors onto the separating hyperplane intersect at a single Radon point.

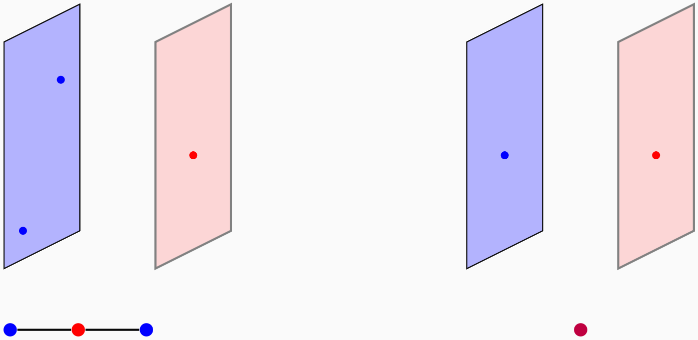
Theorem

Suppose $X \subseteq \mathbb{R}^n$ is in strong general position, and that X is equipped with linearly-separable labels. Then there are at most $n+1$ supporting vectors.

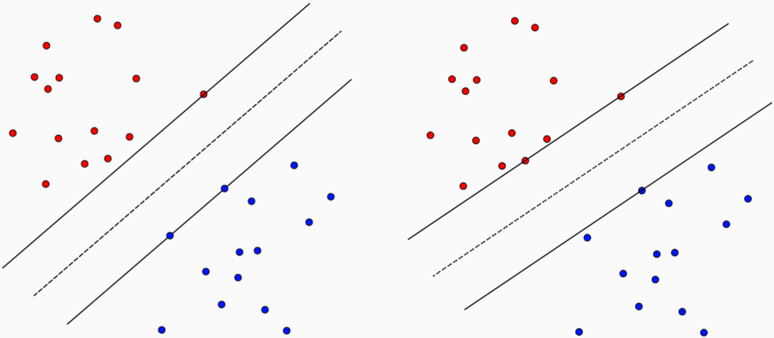
Radon's theorem and SVM configurations



Radon's theorem and SVM configurations



Radon's theorem and SVM configurations



Radon's theorem and SVM configurations

What is preserved under small perturbations of the data points?

Lemma

If $X \subseteq \mathbb{R}^n$ is a set of linearly separable labeled points with positive margin, then there exists an $\varepsilon > 0$ such that upon perturbing any point by at most ε, X remains linearly separable.

Radon's theorem and SVM configurations

What is preserved under small perturbations of the data points?

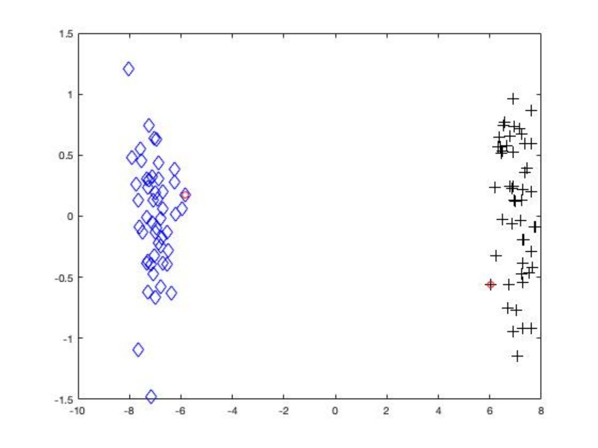
Lemma

If $X \subseteq \mathbb{R}^n$ is a set of linearly separable labeled points with positive margin, then there exists an $\varepsilon > 0$ such that upon perturbing any point by at most ε, X remains linearly separable.

Conjecture

Let $X \subseteq \mathbb{R}^n$ be a set of linearly separable labeled points in strong general position. Let $\varepsilon_0 > 0$ be the minimum distance between any two distinct points in X. Then there exists an $\varepsilon > 0$ with $\varepsilon < \frac{\varepsilon_0}{2}$ such that if each point is perturbed by at most ε, then the set of supporting vectors remains unchanged.

This means figures such as these cannot happen when the data is in strong general position.



Extensions

- Soft margin support vector machines
- Kernel method for linearly inseparable data
- Spherical and ellipsoidal support vector machines
- Probability of obtaining support vector configuration over another
- Firey's dice problem

Acknowledgements

I would like to thank Dr. Henry Adams, Dr. Patrick Shipman, Dr. Anders Fremstad, and Dr. Elly Farnell. I would also like to thank you all for attending!

References I

A.K.V. A. R. Vasishtha, *Matrices*, Krishna Prakashan, 1991.

Lenore Blum, Felipe Cucker, Michael Shub, and Steve Smale, *Complexity and real computation*, Springer Science & Business Media, 2012.

Ake Bjorck, *Numerical methods in matrix computations*, vol. 59, Springer.

Kenneth L Clarkson, David Eppstein, Gary L Miller, Carl Sturtivant, and Shang-Hua Teng, *Approximating center points with iterative radon points*, International Journal of Computational Geometry & Applications **6** (1996), no. 03, 357–377.

D.S. Dummit and R.M. Foote, *Abstract algebra*, Wiley, 2004.

Herbert Edelsbrunner and Ernst Peter Mücke, *Simulation of simplicity: A technique to cope with degenerate cases in geometric algorithms*, ACM Transactions on Graphics (TOG) **9** (1990), no. 1, 66–104.

Martin Fink, John Hershberger, Nirman Kumar, and Subhash Suri, *Separability and convexity of probabilistic point sets*.

William J Firey, *An integral-geometric meaning for lower order area functions of convex bodies*, Mathematatika **19** (1972), no. 2, 205–212.

______, *Kinematic measures for sets of support figures*, Mathematika **21** (1974), no. 2, 270–281.

References II

Robin Hartshorne, *Algebraic geometry*, vol. 52, Springer Science & Business Media, 2013.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola, *Kernel methods in machine learning*, The annals of statistics (2008), 1171–1220.

Vojislav Kecman, *Support vector machines—An introduction*, Support vector machines: Theory and applications, Springer, 2005, pp. 1–47.

Gregor Kemper, *A course in commutative algebra*, vol. 256, Springer Science & Business Media, 2010.

Jeremy Kun, *Duality for the SVM*.

Lynn Harold Loomis and Shlomo Sternberg, *Advanced calculus*, World Scientific, 1968.

Jiří Matoušek, *Lectures on discrete geometry*, Springer-Verlag, Berlin, Heidelberg, 2002.

——————, *Using the Borsuk–Ulam theorem: Lectures on topological methods in combinatorics and geometry*, Springer, 2003.

P McMullen, *A dice probability problem*, Mathematika **21** (1974), no. 2, 193–198.

Yunqian Ma and Guodong Guo, *Support vector machines applications*, Springer Publishing Company, Incorporated, 2014.

- K-R Muller, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, and Bernhard Scholkopf, *An introduction to kernel-based learning algorithms*, IEEE transactions on neural networks **12** (2001), no. 2, 181–201.
- BB Peterson, *The geometry of Radon's theorem*, American Mathematical Monthly (1972), 949–963.
- John Platt, *Sequential minimal optimization: A fast algorithm for training support vector machines*, Tech. report, Microsoft, April 1998.
- Luis David Garcia Puente, *Chapter 1: Varieties*, `http://www.shsu.edu/~ldg005/data/689/L1.pdf`.
- Johann Radon, *Mengen konvexer körper, die einen gemeinsamen punkt enthalten*, Mathematische Annalen **83** (1921), no. 1-2, 113–115.
- George E Sakr and Imad H Elhajj, *VC-based confidence and credibility for support vector machines*, Soft Computing **20** (2016), no. 1, 133–147.
- Alex J Smola and Bernhard Schölkopf, *A tutorial on support vector regression*, Statistics and computing **14** (2004), no. 3, 199–222.
- Rolf Schneider and Wolfgang Weil, *Stochastic and integral geometry*, Springer, 2008.
- Lloyd N Trefethen and David Bau III, *Numerical linear algebra*, vol. 50, Siam, 1997.

- Ivor W Tsang, James T Kwok, and Pak-Ming Cheung, *Core vector machines: Fast svm training on very large data sets*, Journal of Machine Learning Research **6** (2005), 363–392.
- Helge Tverberg, *A generalization of radon's theorem*, Journal of the London Mathematical Society **1** (1966), no. 1, 123–128.
- Vladimir Vapnik, *Statistical learning theory. 1998*, vol. 3, Wiley, New York, 1998.
- ______, *The nature of statistical learning theory*, Springer science & business media, 2013.
- V. Vapnik and A. Chervonenkis, *On the uniform convergence of relative frequencies of events to their probabilities*, Theory of Probability & Its Applications **16** (1971), no. 2, 264–280.
- Chih-Chia Yao, *Utilizing ellipsoid on support vector machines*, Machine Learning and Cybernetics, 2008 International Conference on, vol. 6, IEEE, 2008, pp. 3373–3378.