On Support Vector Machines

CS189/289A: Introduction to Machine Learning

Stella Yu

UC Berkeley

24 October 2017

Key Concepts and Different Views of SVM

- Decision function vs. generative/discriminative models

- View 1: Max margin principle and primal SVM

- View 2: Slack and Tikhonov regularization learning

- View 3: Convex hulls and dual SVM

Classification: Predict *Default* from (Balance,Income)

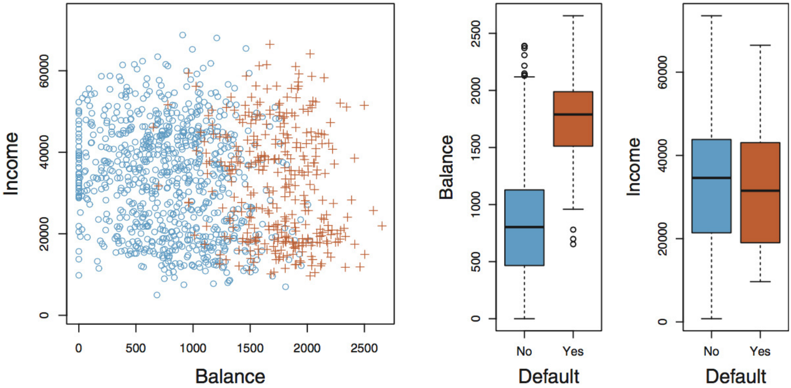
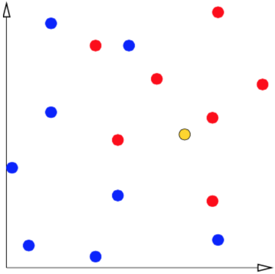


FIGURE 4.1. *The* Default *data set.* Left: *The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.* Center: *Boxplots of* balance *as a function of* default *status.* Right: *Boxplots of* income *as a function of* default *status.*

The Problem of Learning A Feature Classifier



- We have paired training data, feature x_i and class label y_i, $i = 1, \ldots, n$.
- There is uncertainty in relating input feature x to output class label y, because the feature may not uniquely specify the class.
- What is the best guess of y for a new point x?

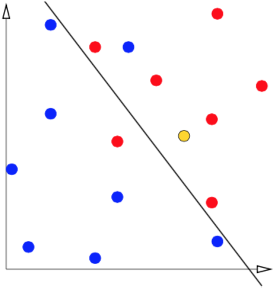
Three Types of Machine Learning Models

1. **Probablistic models** view learning as a process of reducing uncertainty, modeled by means of probability distributions.

$$\text{MLE:} \qquad y = \arg\max_y P(x|y) \tag{1}$$

$$\text{MAP:} \qquad y = \arg\max_y P(y|x) \tag{2}$$

2. **Geometric models** use geometrical intuitions such as separating planes, linear transformations, distance metrics etc.



3. **Logical models** are defined by easily interpretable rules.

Ways of Building A Feature Classification Model

1. **Generative:** Model the class conditional $P(x|y)$ and the prior $P(y)$, obtain the posterior $P(y|x)$ using the Bayes Theorem.

2. **Discriminative:** Model the posterior $P(y|x)$ directly.

3. **Decision boundaries:** Model $f(x) \rightarrow y$.

   We don't care about posteriors. We just want the decision rule to choose the class based on the feature vector. We can be in error but never in doubt. Here we cannot detect outliers.

   Vapnik: When you are trying to solve a hard problem, don't solve an intermediate problem which is even harder.

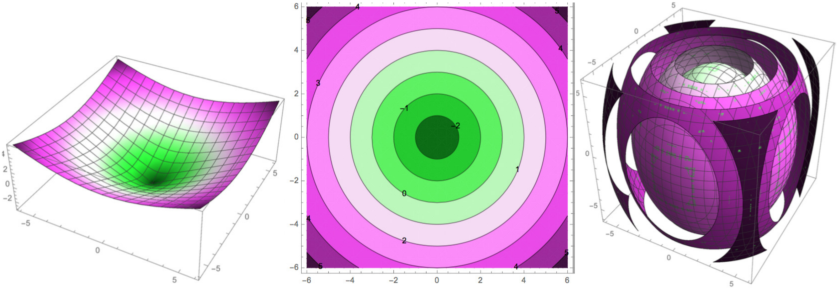
Decision Function and Decision Boundary

- Decision / predictor / discriminant function $f(x)$ maps a point x to a scalar such that

$$f(x) > 0, \qquad \text{if } x \in \text{ class } C \tag{3}$$

$$f(x) \leq 0, \qquad \text{if } x \notin \text{ class } C \tag{4}$$

- The decision boundary is $\{x : f(x) = 0\}$, the set of all points where the decision function is zero, usually a $d-1$ dimensional surface in the d dimensional feature space.
- Examples of isosurface of f for isovalue 0, ±1, ...



Geometric Interpretation of Linear Functions

- Linear function defined by inner product:

$$f(x) = w \cdot x - t \tag{5}$$

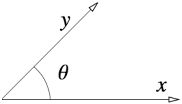
- The length of a vector:

$$|x| = \sqrt{x \cdot x} = \sqrt{x_1^2 + x_2^2 + \ldots + x_d^2} \tag{6}$$

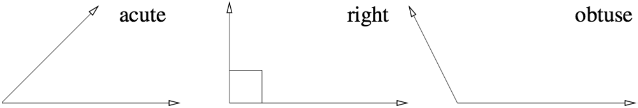
- The normalized vector indicates the direction of a vector:

$$\left| \frac{x}{|x|} \right| = 1 \tag{7}$$

- The angles between two vectors:



$$\cos\theta = \frac{x \cdot y}{|x|\,|y|} = \underbrace{\frac{x}{|x|}}_{\text{length } 1} \cdot \underbrace{\frac{y}{|y|}}_{\text{length } 1}$$



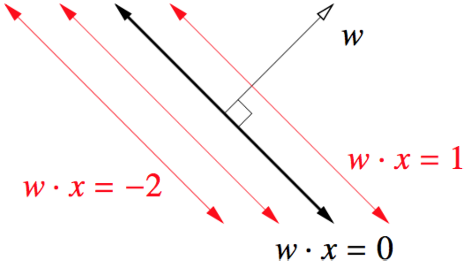
$\cos\theta > 0$ $\cos\theta = 0$ $\cos\theta < 0$

Geometric Interpretation of Linear Functions

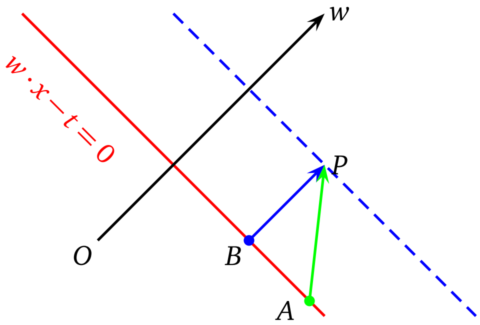
- The linear decision boundary is a hyperplane:

$$H = \{x : f(x) = 0, \text{ or } w \cdot x = t\} \tag{8}$$

- In 2D, H is a line. In 3D, H is a plane. In d dimensions, H is a hyperplane – a flat, infinite thing with dimension $d-1$. It divides the d-dimensional space into two halves.
- The normal vector of H is w, since for any two points x and z that lie on H, $w \cdot (x-z) = t - t = 0$.



Distance From A Point to A Hyperplane



$$\text{hyperplane } H: \quad w \cdot x - t = 0 \tag{9}$$

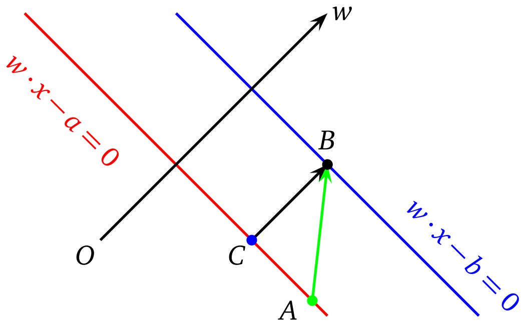
$$\text{arbitrary point } P: \quad z \tag{10}$$

$$\text{any point } A \text{ on } H: \quad x, \quad \text{where } w \cdot x - t = 0 \tag{11}$$

$$\text{distance from } P \text{ to } H: \quad \text{distance}(P, H) = \text{Project}_w PA \tag{12}$$

$$= \frac{|w \cdot (z - x)|}{|w|} = \frac{|w \cdot z - t|}{|w|} \tag{13}$$

Distance Between Two Parallel Lines



$$\text{hyperplane } H_1: \quad w \cdot x - a = 0 \tag{14}$$

$$\text{hyperplane } H_2: \quad w \cdot x - b = 0 \tag{15}$$

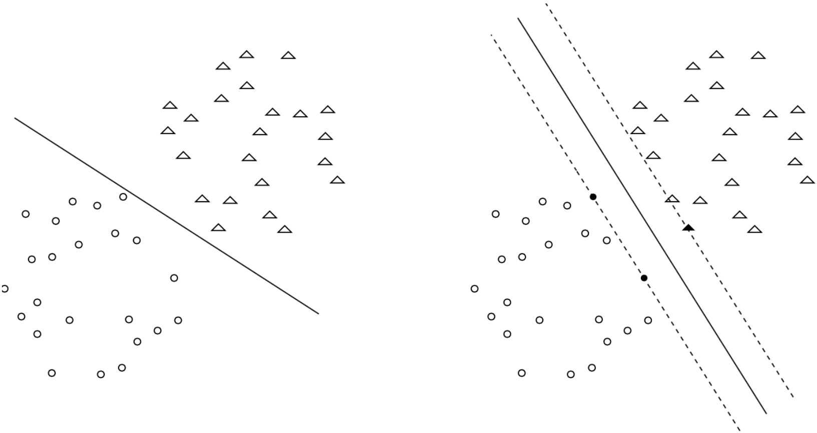
$$\text{any point } A \text{ on } H_1: \quad x, \quad \text{where } w \cdot x - a = 0 \tag{16}$$

$$\text{any point } B \text{ on } H_2: \quad z, \quad \text{where } w \cdot z - b = 0 \tag{17}$$

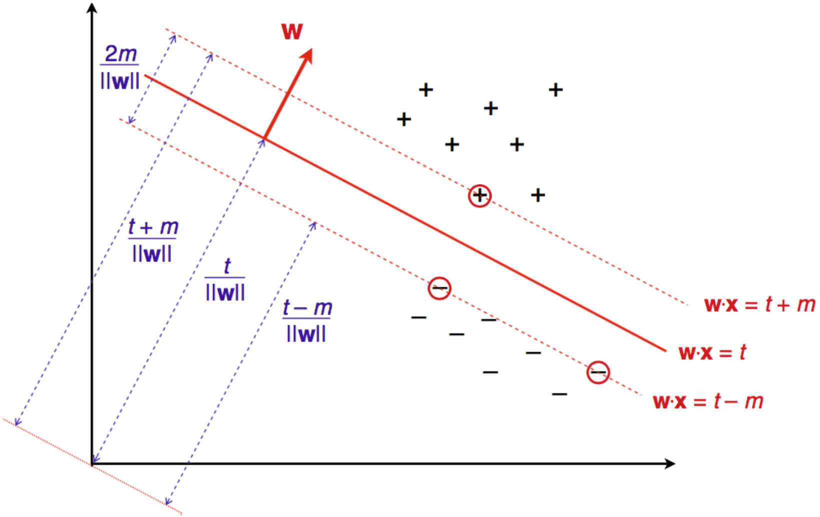
$$\text{distance}(H_1, H_2): \quad = \text{Project}_w AB \tag{18}$$

$$= \frac{|w \cdot (z - x)|}{|w|} = \frac{|a - b|}{|w|} \tag{19}$$

Maximum Margin Principle: Avoid Overfitting



The Geometry of A Support Vector Machine



w
$\frac{2m}{||\mathbf{w}||}$
$\frac{t+m}{||\mathbf{w}||}$
$\frac{t}{||\mathbf{w}||}$
$\frac{t-m}{||\mathbf{w}||}$
$\mathbf{w}\cdot\mathbf{x} = t + m$
$\mathbf{w}\cdot\mathbf{x} = t$
$\mathbf{w}\cdot\mathbf{x} = t - m$

Maximize the Margin in Linearly Separable Cases

- We want to maximize the margin $\frac{2m}{|w|}$, the distance between two lines $w \cdot x = t - m$ and $w \cdot x = t + m$.

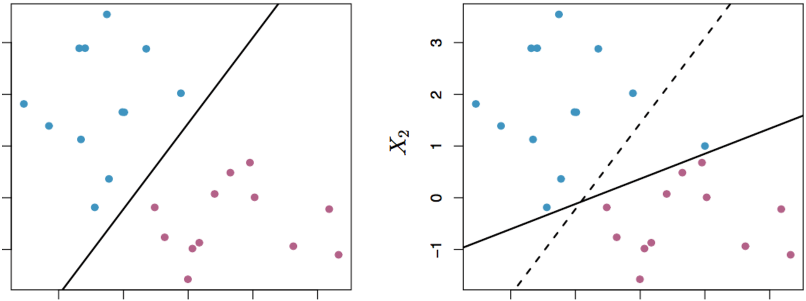
- Since we are free to rescale $t, |w|, m$, we could choose $m = 1$.

- Maximizing the margin then corresponds to minimizing $|w|$ or, more conveniently, $\frac{1}{2}|w|^2$, provided of course that none of the training points fall inside the margin.

$$\min_{w,t} \quad \frac{1}{2}|w|^2 \tag{20}$$

$$\text{s. t.} \quad y_i(w \cdot x_i - t) \geq 1, \quad i = 1, 2, \ldots, n \tag{21}$$

Hard-Margin SVMs Are Brittle



1. Hard-margin SVMs are sensitive to outliers.

2. Hard-margin SVMs fail if data are not linearly separable.

Soft-Margin SVM: Build in Slacks

- Allow some points to violate the margin, with slack variables.

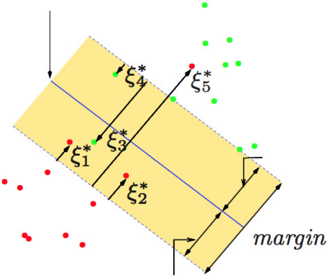
$$y_i(x_i \cdot w - t) \geq 1 - \xi_i \tag{22}$$

- Points that don't violate the margin are treated the same:

$$y_i(x_i \cdot w - t) \geq 1 \quad \Rightarrow \quad \xi_i = 0 \tag{23}$$

- Points that violate the margin have non-zero slacks – lower the margin standards for those inside the margin or even at the wrong side of the decision boundary:

$$y_i(x_i \cdot w - t) \geq 1 - \xi \quad \Rightarrow \quad \xi_i > 0 \tag{24}$$



Soft-Margin SVM

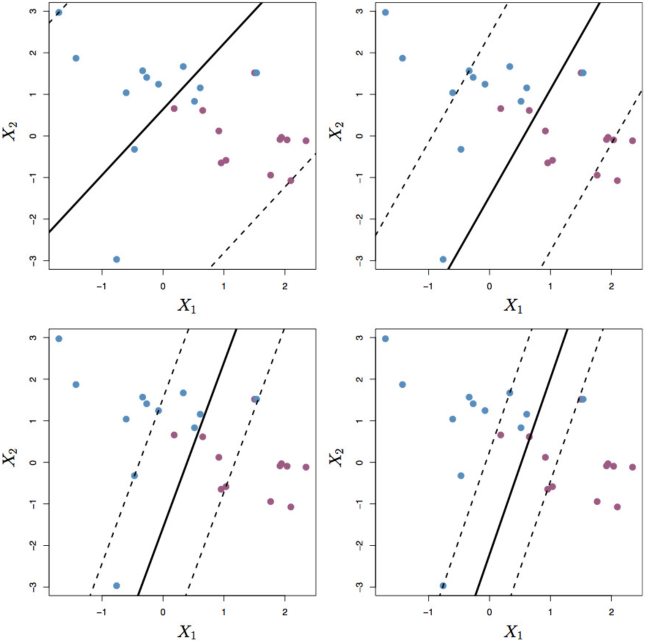
- To prevent abuse of slack, we add a loss term on slacks.

$$
\begin{aligned}
\min_{w,t,\xi} \quad & \frac{1}{2}|w|^2 + C\sum_{i=1}^{n} \xi_i && && (25)\\
\text{s. t.} \quad & y_i(w \cdot x_i - t) \geq 1 - \xi_i, && i = 1, 2, \ldots, n && (26)\\
& \xi_i \geq 0, && i = 1, 2, \ldots, n && (27)
\end{aligned}
$$

- A quadratic program in $d + 1 + n$ dimensions, $2n$ constraints
- C is a scalar hyperparameter (use validation!) that trades off:

	small C	big C
desire	maximize margin	keep slacks small
danger	underfitting	overfitting
outliers	less sensitive	very sensitive
boundary	more flat	more sinuous

Bigger C, Less Violation, Smaller Margin



Classification by Tikhonov Regularization Learning

- Decision function $f(x)$ for binary classification

$$y = \mathrm{sign}(f(x)) \tag{28}$$

- Tikhonov Regularization Learning:

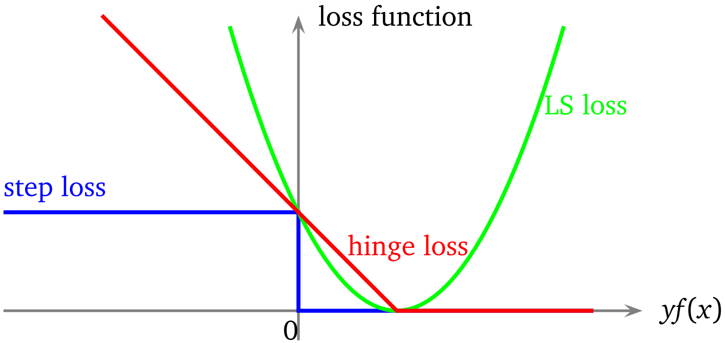
$$\min_{f} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \|f\|^2 \tag{29}$$

- Step loss function $L(y, f(x))$ penalizes #misclassification:

$$L(y_i, f(x_i)) = \begin{cases} 1, & y_i f(x_i) < 0 \\ 0, & y_i f(x_i) \geq 0 \end{cases} \tag{30}$$

Cons: Neither convex nor differentiable.

SVM = Tikhonov Regularization with Hinge Loss



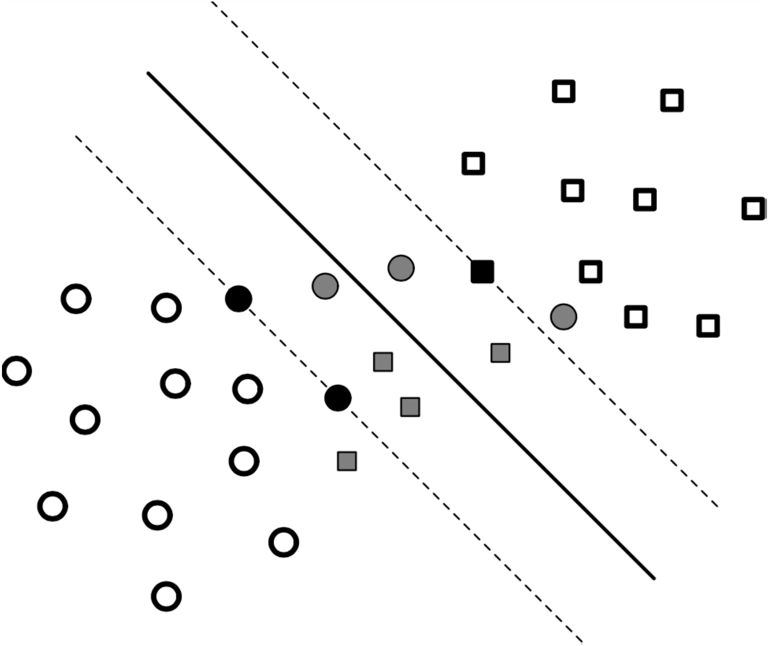
- Hinge Loss is convex, although not differentiable:

$$\min_f L(f) = \frac{1}{n}\sum_{i=1}^{n} \max(1 - y_i f(x_i), 0) + \lambda\|f\|^2 \quad (31)$$

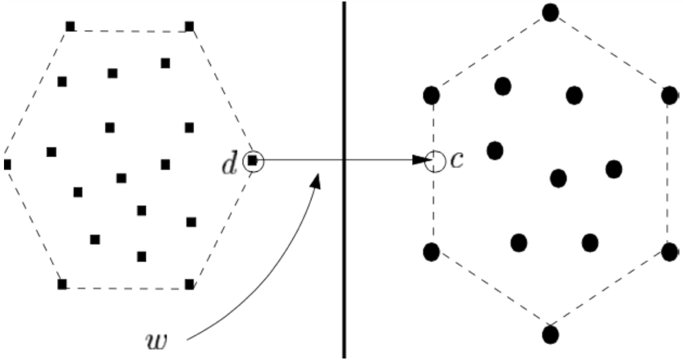
$$\max(1 - y_i f(x_i), 0) \equiv \xi_i \quad (32)$$

$$\lambda \equiv \frac{1}{C} \quad (33)$$

Support Vectors for the Optimal SVM Classifier

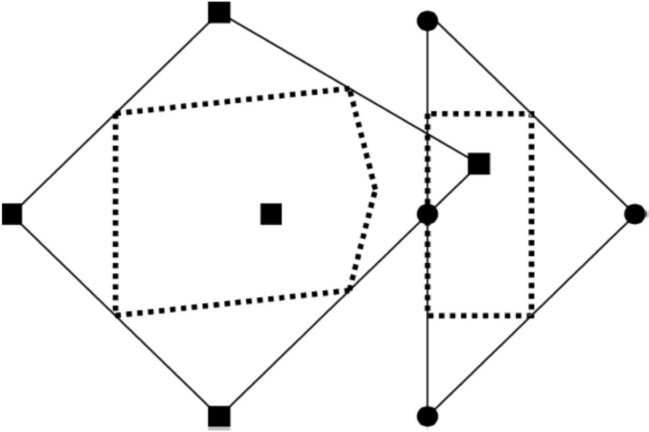


Interpretation from the Dual SVM: Hard Margin



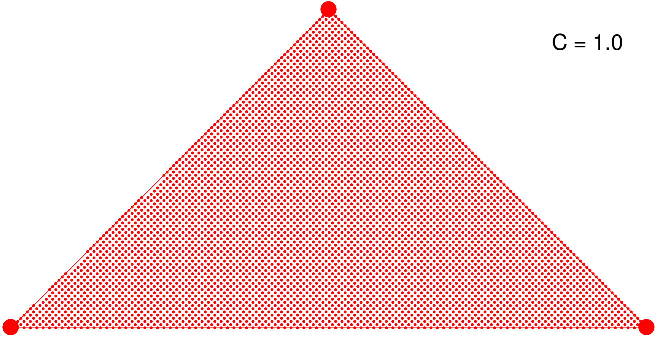
- Let the primal SVM be defined by two parallel bounding planes, parameterized by normal w and two offsets a, b.
- The primal SVM finds two parallel bounding planes of max separation between point clouds. The dual SVM finds the closest points between two convex hulls.
- *Duality and Geometry in SVM Classifiers, Kristin P. Bennett and Erin J. Bredensteiner, ICML, 2000*

Interpretation from the Dual SVM: Soft Margin



- When two convex hulls intersect, the SVM contracts or reduces the convex hulls by putting an upperbound on the multiplier in the convex combination for each point.
- *Power SVM: Generalization with Exemplar Classification Uncertainty, Weiyu Zhang, Stella X. Yu,Shang-Hua Teng, CVPR 2012*

Convex Hull of A Triangle

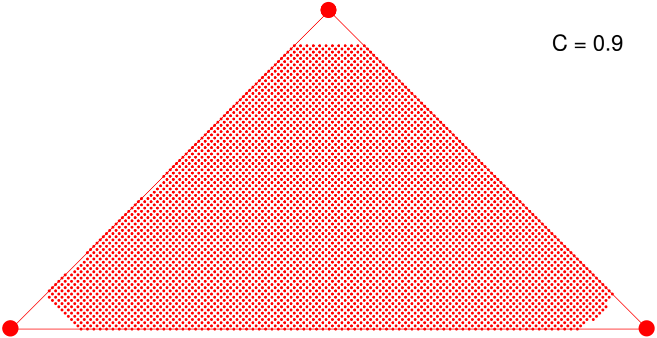


convex combination: $$z = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 \tag{34}$$

weight normalization: $$\alpha_1 + \alpha_2 + \alpha_3 = 1 \tag{35}$$

nonnegative weights: $$\alpha_i \geq 0, \quad i = 1, 2, 3. \tag{36}$$

Reduced Convex Hull of A Triangle

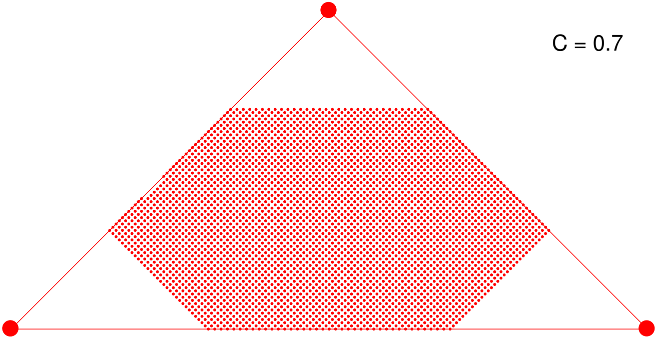


convex combination: $$z = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 \qquad (37)$$

weight normalization: $$\alpha_1 + \alpha_2 + \alpha_3 = 1 \qquad (38)$$

weight range: $$0 \le \alpha_i \le C = 0.9, \quad i = 1, 2, 3. \qquad (39)$$

Reduced Convex Hull of A Triangle

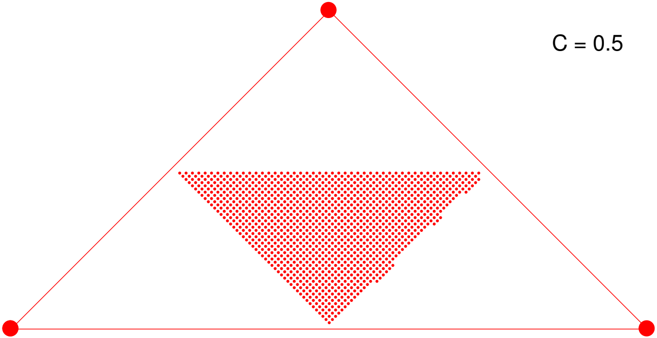


convex combination: $$z = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 \tag{40}$$

weight normalization: $$\alpha_1 + \alpha_2 + \alpha_3 = 1 \tag{41}$$

weight range: $$0 \leq \alpha_i \leq C = 0.7, \quad i = 1, 2, 3. \tag{42}$$

Reduced Convex Hull of A Triangle

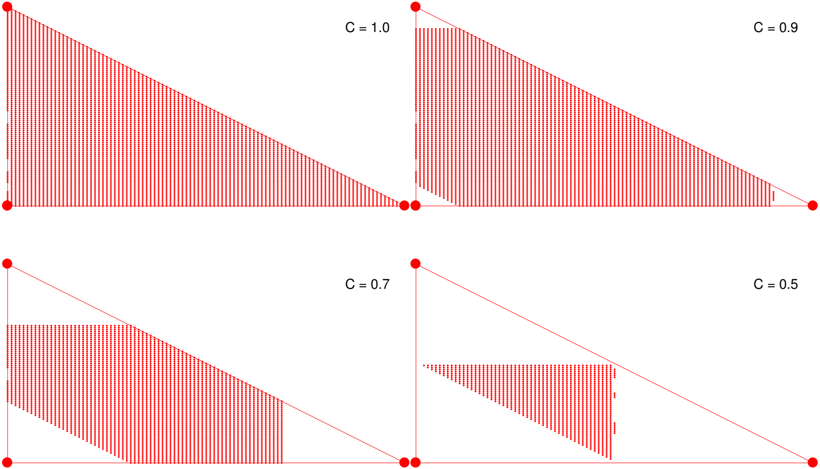


convex combination: $$z = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 \tag{43}$$

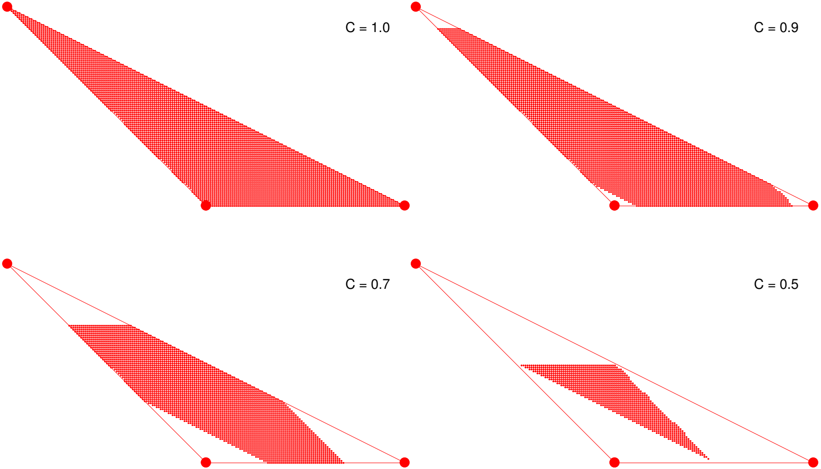
weight normalization: $$\alpha_1 + \alpha_2 + \alpha_3 = 1 \tag{44}$$

weight range: $$0 \leq \alpha_i \leq C = 0.5, \quad i = 1, 2, 3. \tag{45}$$

Reduced Convex Hulls of Triangles



Reduced Convex Hulls of Triangles


C = 1.0
C = 0.9
C = 0.7
C = 0.5