CPSC 340:
Machine Learning and Data Mining

Density-Based Clustering

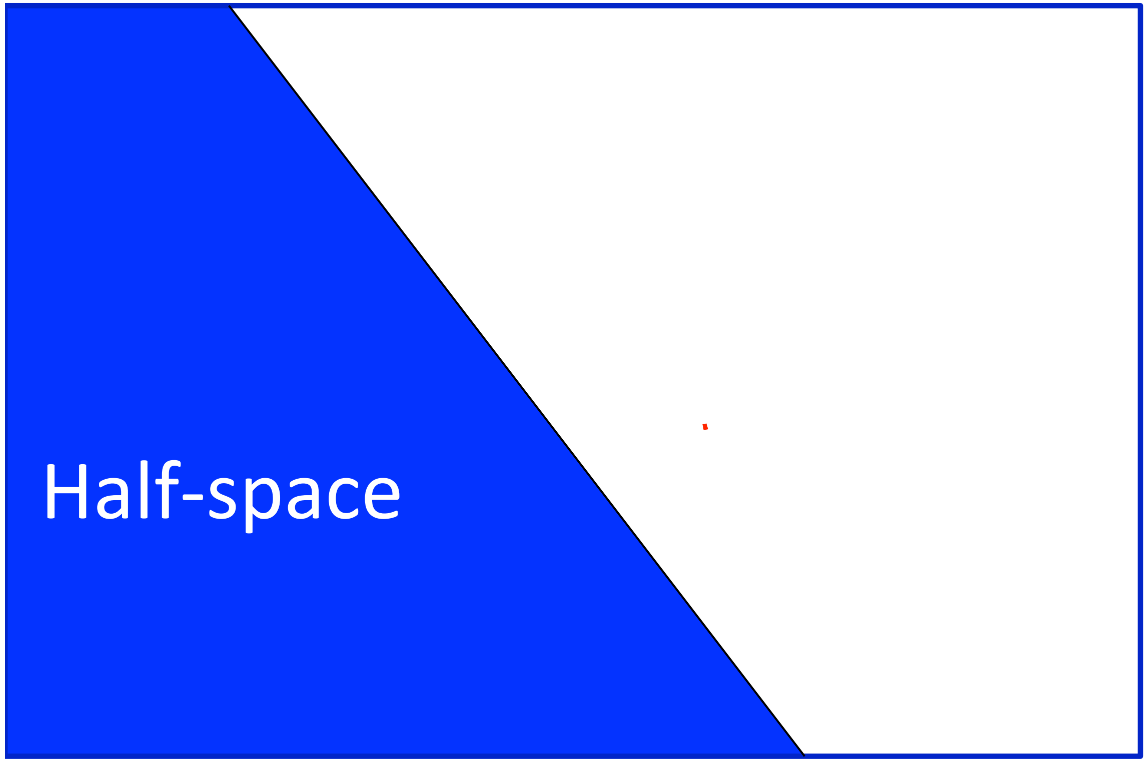
BONUS SLIDES

Shape of K-Means Clusters

- K-means clusters are formed by the intersection of half-spaces.

Half-space is a set of points satisfying a linear inequality, like $\sum_{j=1}^{d} a_j x_j \leq b$

$a^T x$

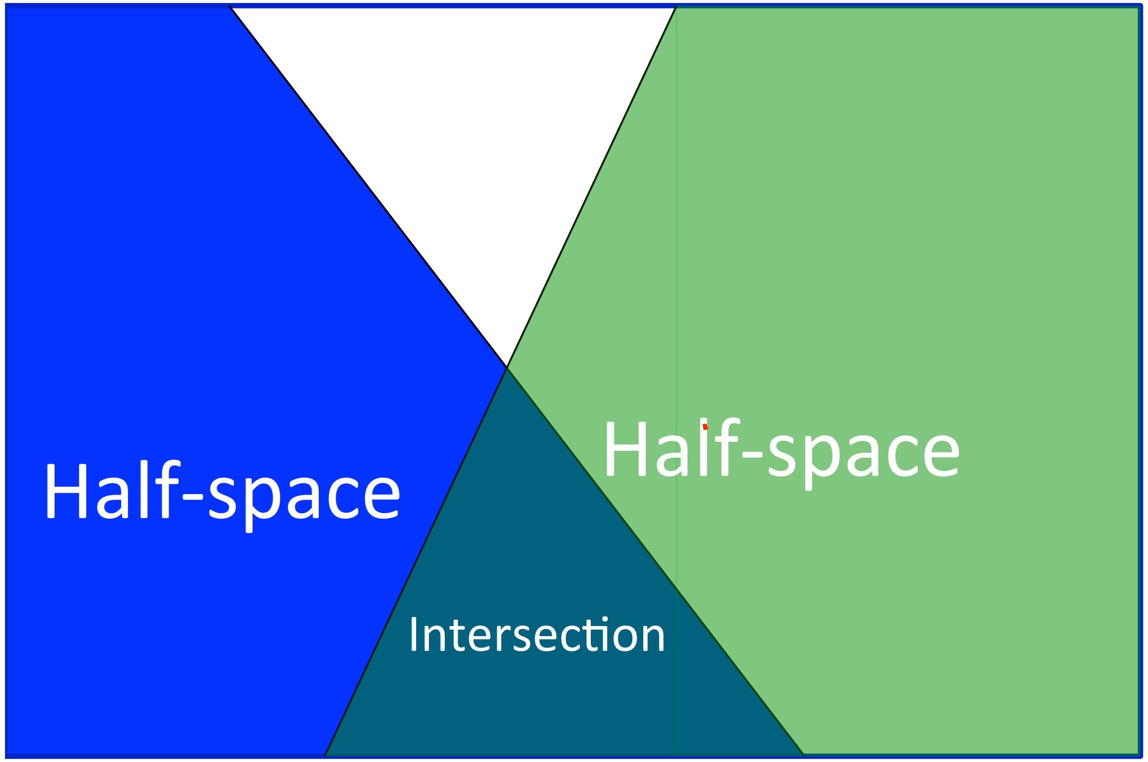


Shape of K-Means Clusters

- K-means clusters are formed by the intersection of half-spaces.

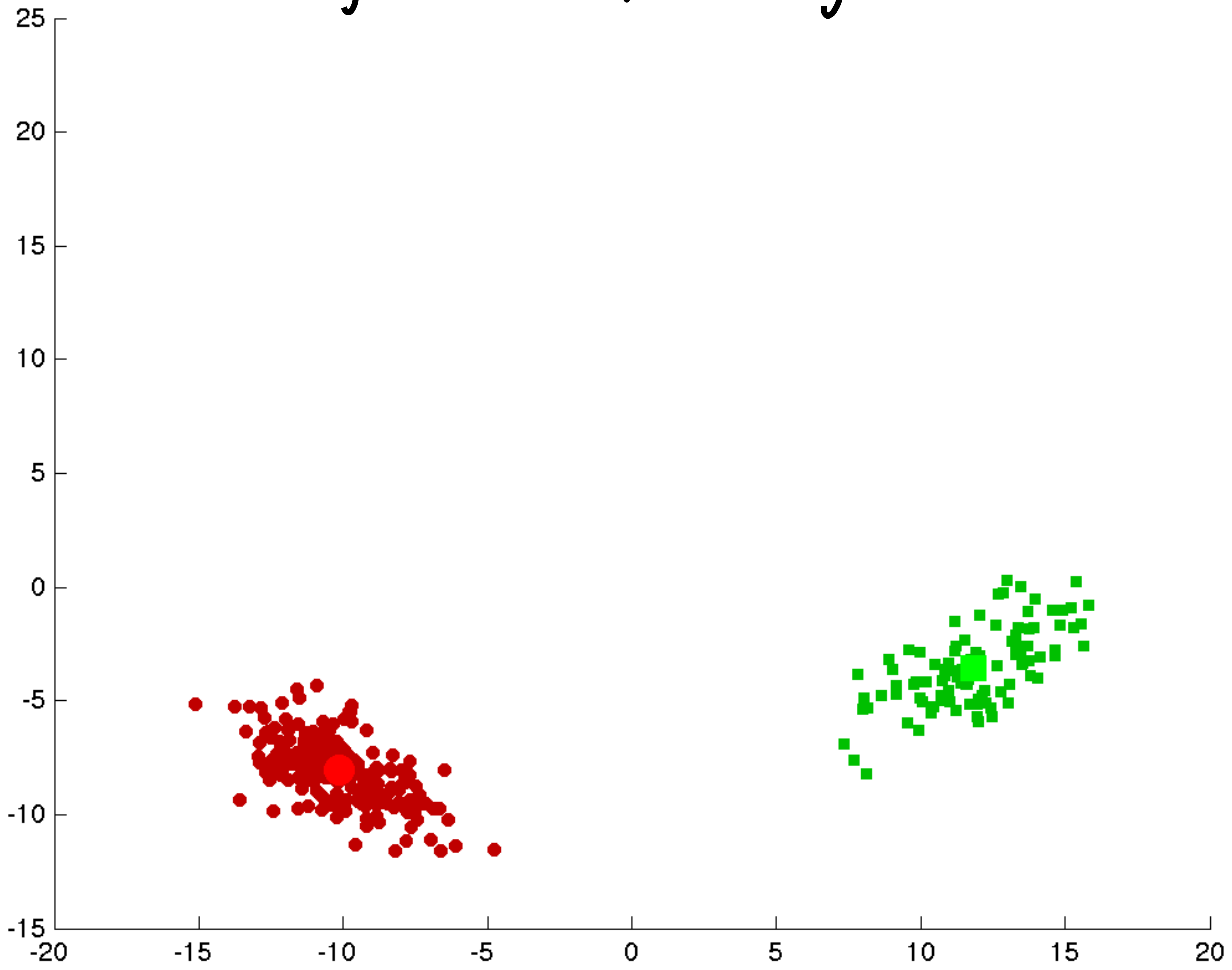
Half-space is a Set of points satifying a linear inequality, like $\sum_{j=1}^{d} a_j x_j \leq b$

'x'



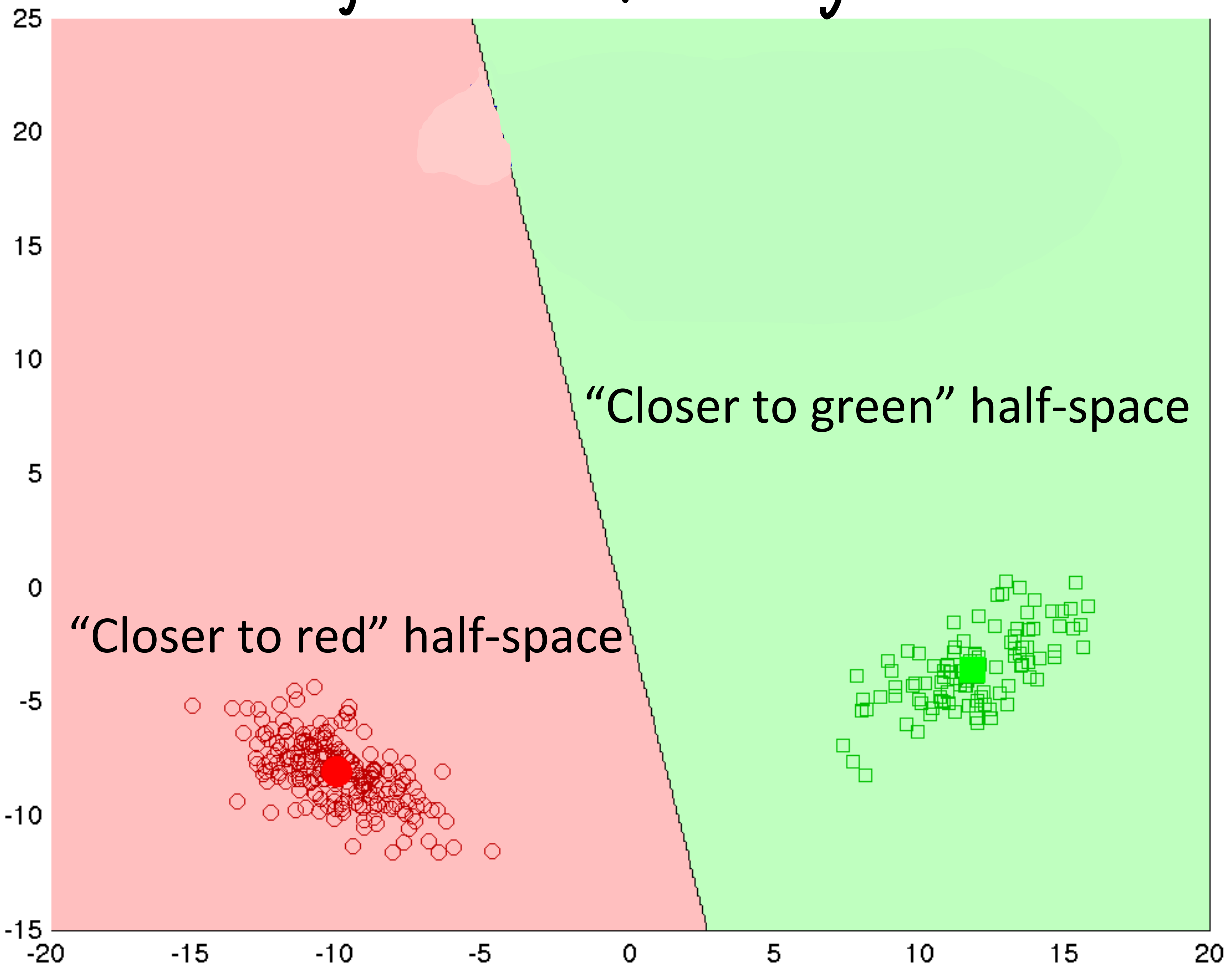
Shape of K-Means Clusters

Which regions are put in green cluster?



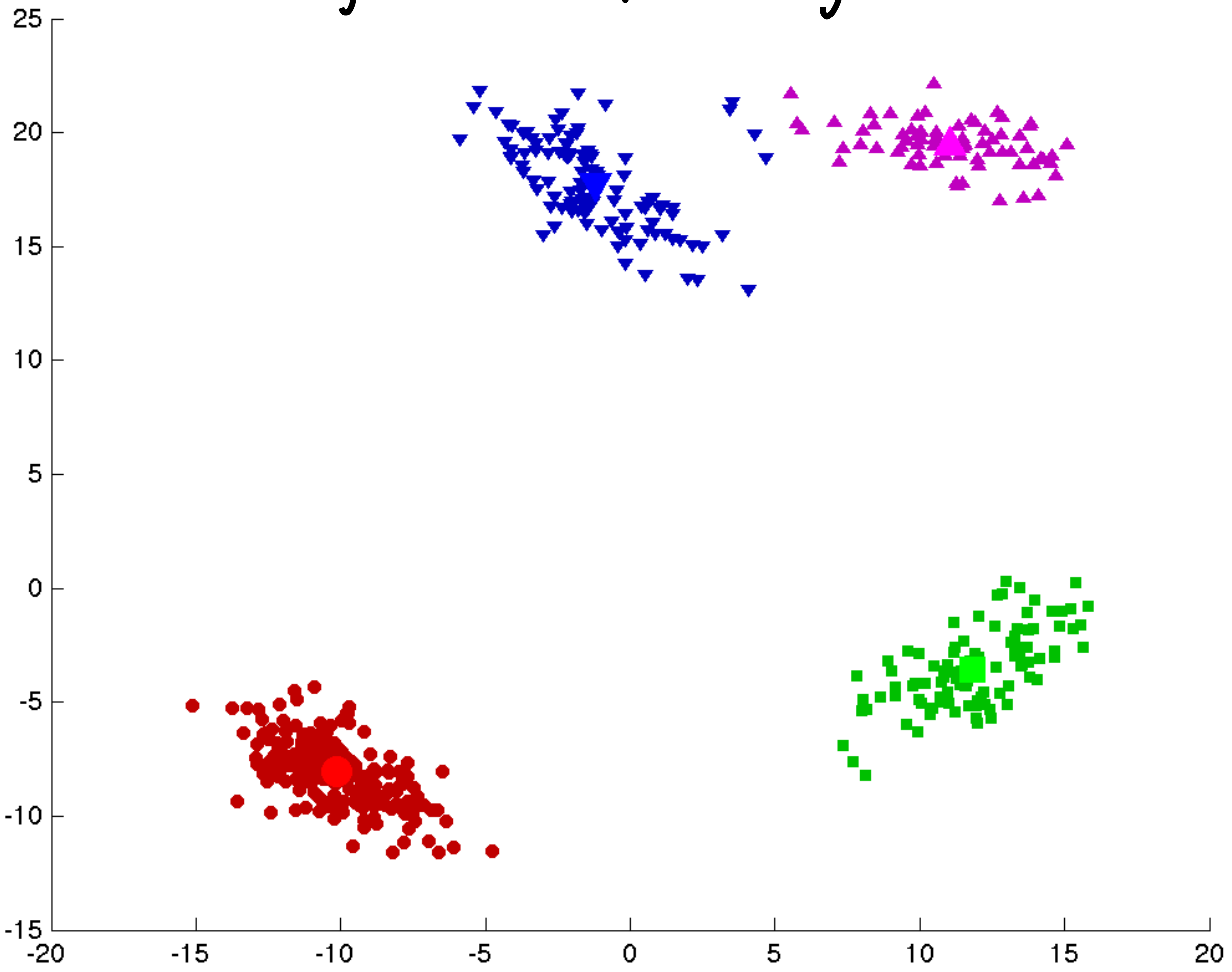
Shape of K-Means Clusters

Which regions are put in green cluster?



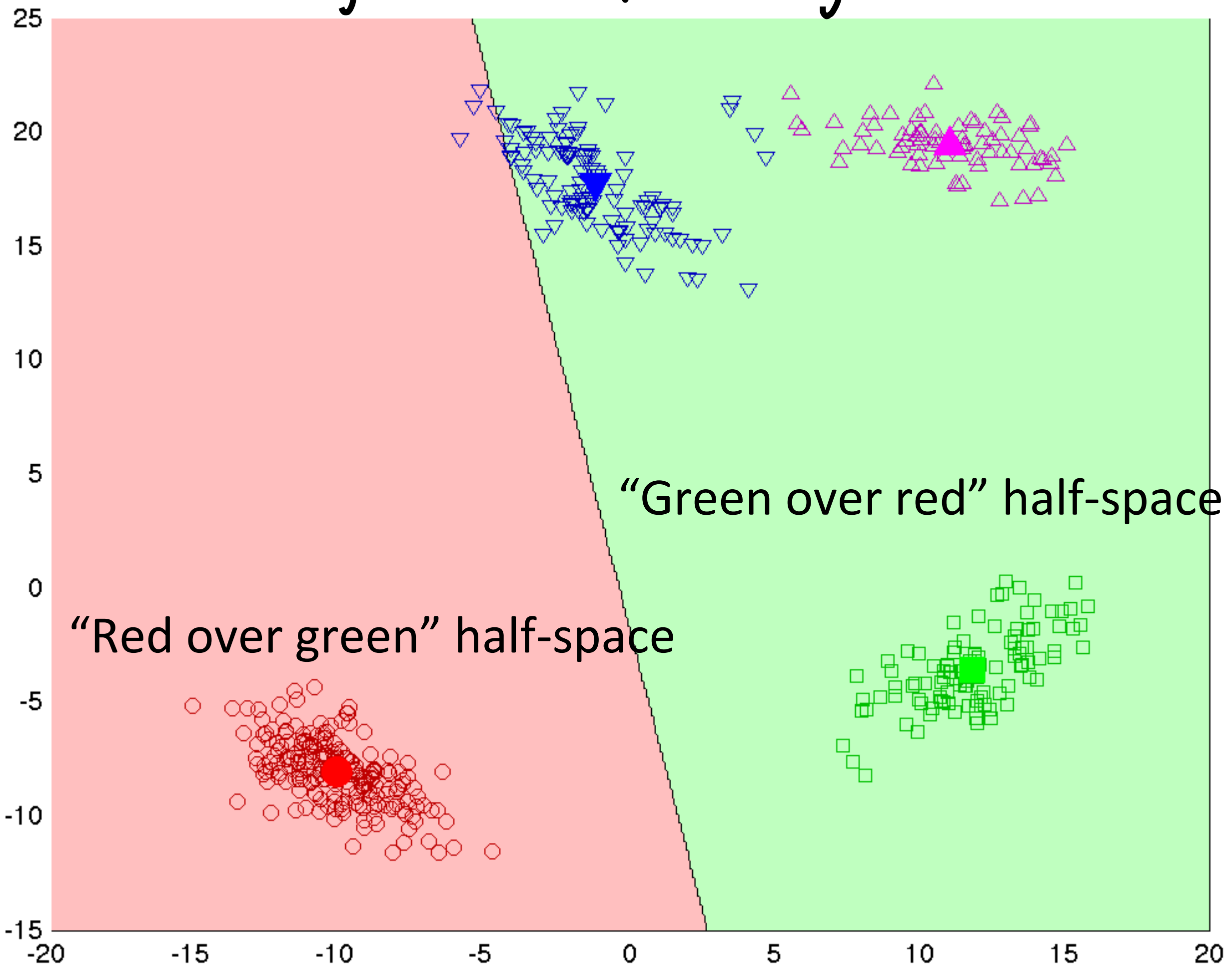
Shape of K-Means Clusters

Which regions are put in green cluster?



Shape of K-Means Clusters

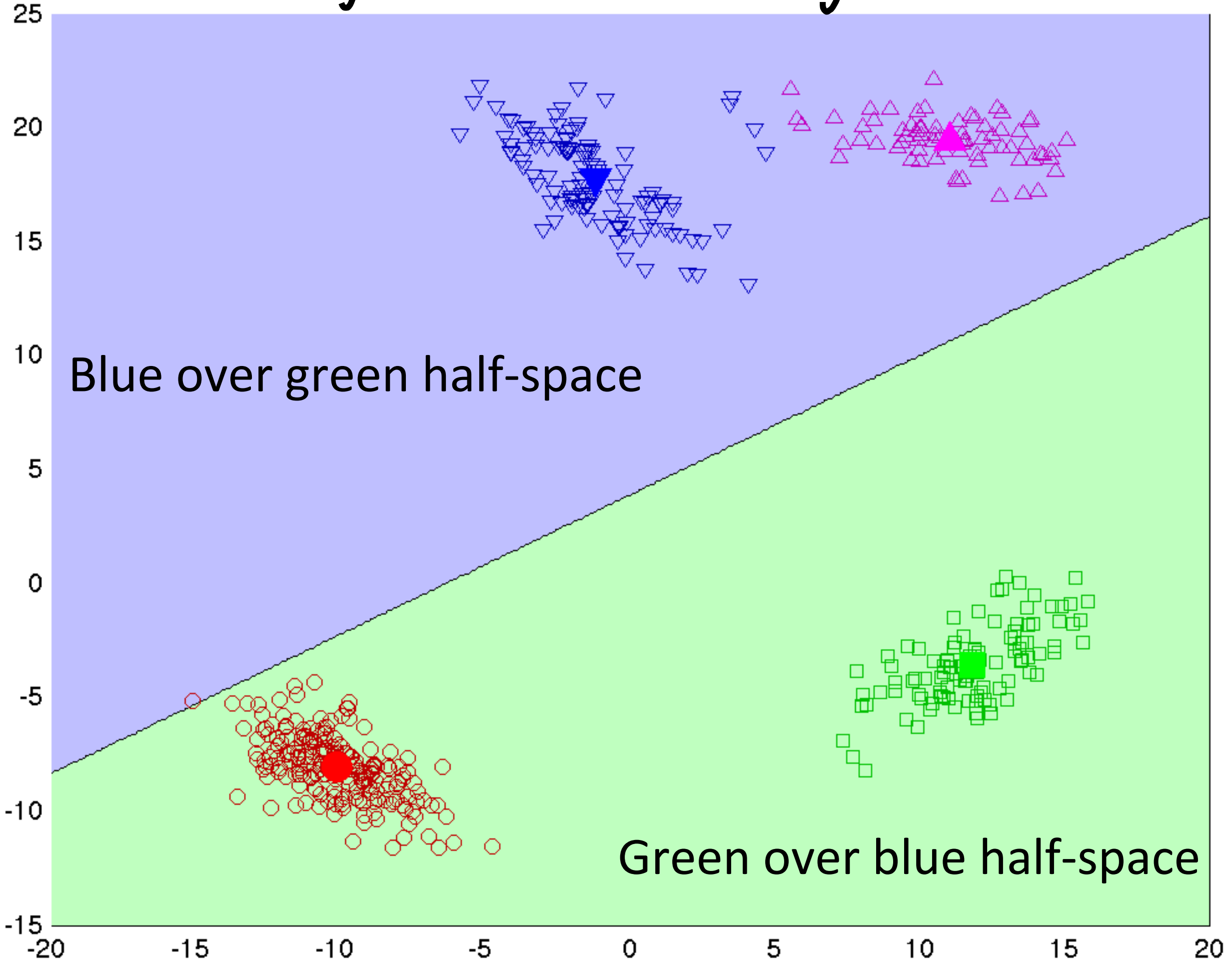
Which regions are put in green cluster?



Red vs. green decision stays the same with more clusters.

Shape of K-Means Clusters

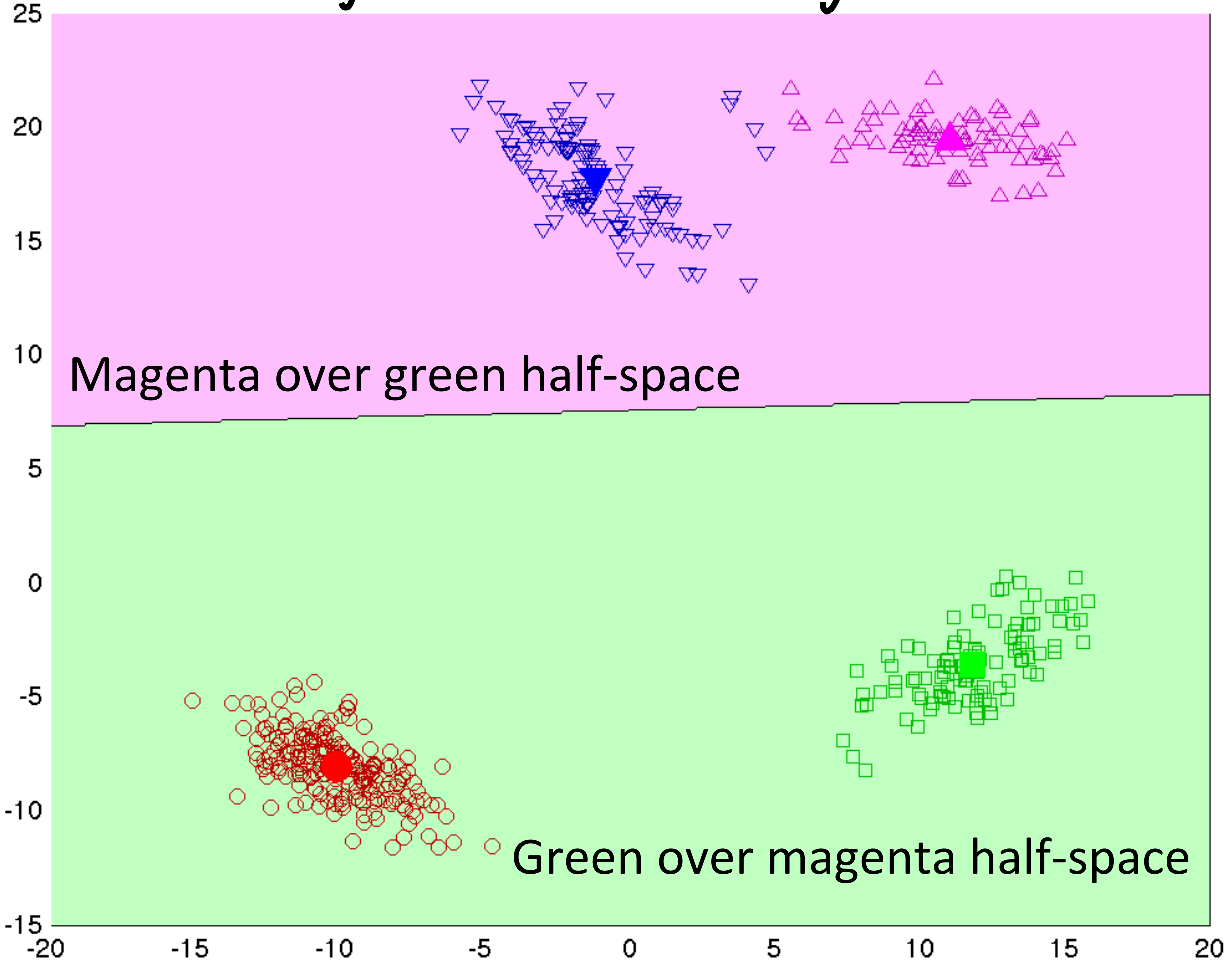
Which regions are put in green cluster?



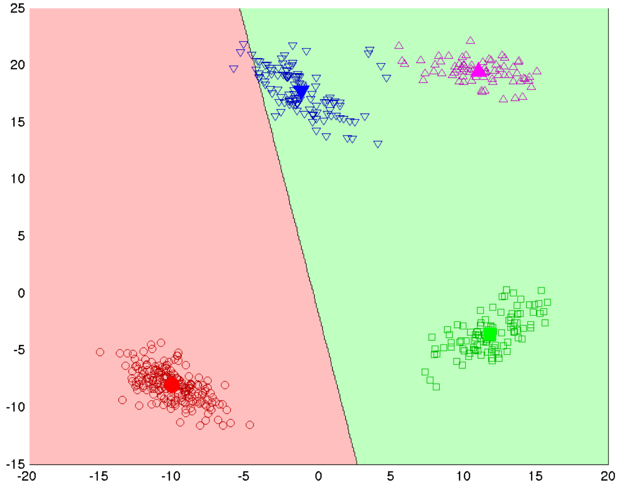
Blue vs. green decision defines different half-spaces.

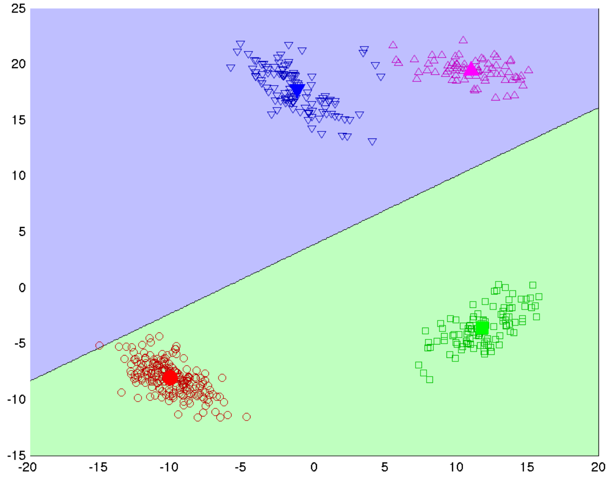
Shape of K-Means Clusters

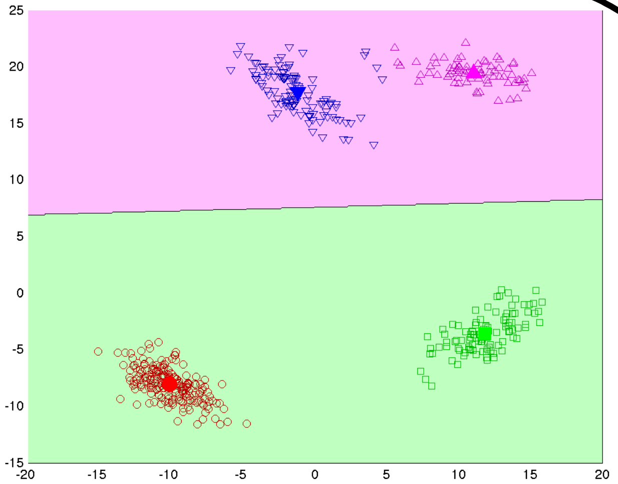
Which regions are put in green cluster?



Shape of K-Means Clusters

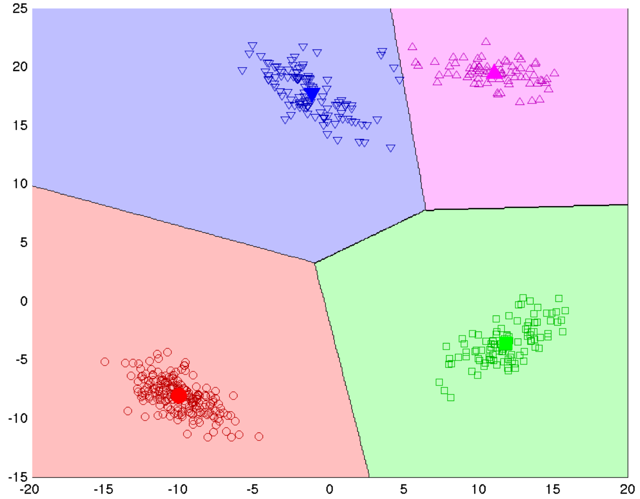






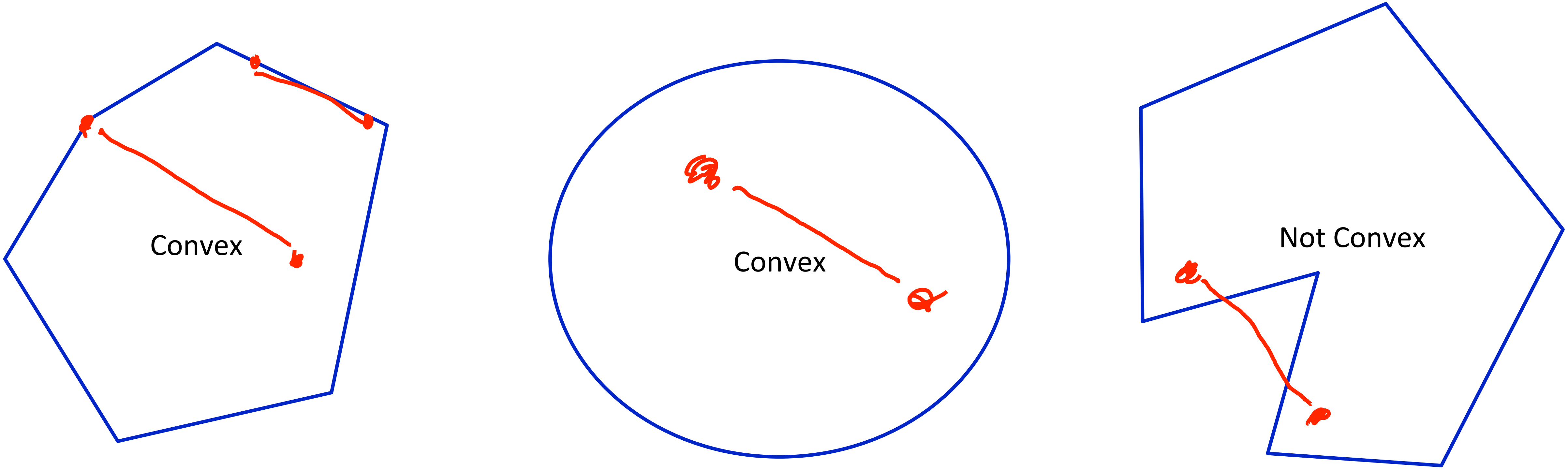
Green "cluster" is the <u>intersection</u> of these three half-spaces.

Here is what the four clusters look like:



Shape of K-Means Clusters

- Intersection of half-spaces form a convex set:
  - Line between any two points in the set stays in the set.



Bonus Slide: Lp-norms

- The L_1-, L_2-, and L_∞-norms are special cases of Lp-norms:

$$||x||_p = \left(\sum_{j=1}^{d} x_j^p \right)^{1/p}$$

- This gives a norm for any (real-valued) $p \geq 1$.
  - The L_∞-norm is limit as 'p' goes to ∞.

- For $p < 1$, not a norm because triangle inequality not satisfied.

Bonus Slide: Squared/Euclidean-Norm Notation

We're using the following conventions:

The subscript after the norm is used to denote the p-norm, as in these examples:

$\|x\|_2 = \sqrt{\sum_{j=1}^{d} w_j^2}$.
$\|x\|_1 = \sum_{j=1}^{d} |w_j|$.

If the subscript is omitted, we mean the 2-norm:

$\|x\| = \|x\|_2$.

If we want to talk about the *squared* value of the norm we use a superscript of "2":

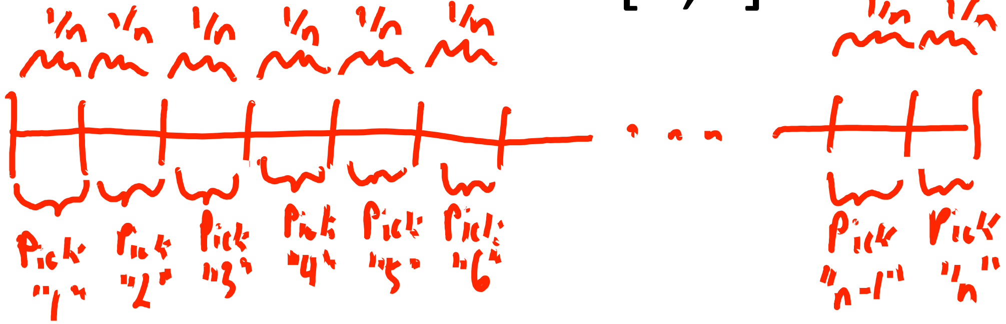
$\|x\|_2^2 = \sum_{j=1}^{d} w_j^2$.
$\|x\|_1^2 = \left(\sum_{j=1}^{d} |w_j|\right)^2$.

If we omit the subscript and have a superscript of "2", we're taking about the squared L2-norm:

$\|x\|^2 = \sum_{j=1}^{d} w_j^2$.

Bonus Slide: Uniform Sampling

- Standard approach to generating a random number from {1,2,…,n}:
  1. Generate a uniform random number ‘u’ in the interval [0,1].
  2. Return the largest index ‘i’ such that $u \leq i/n$.
- Conceptually, this divides interval [0,1] into ‘n’ equal-size pieces:

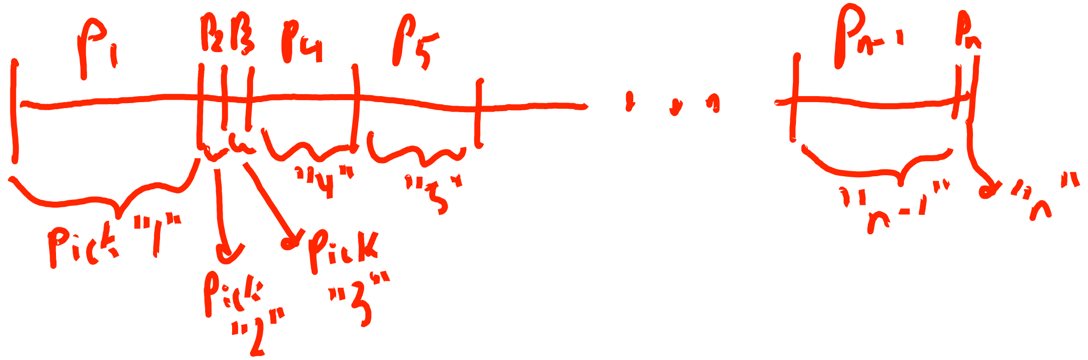


- This assumes $p_i = 1/n$ for all ‘i’.

probability of picking number 'i'.

Bonus Slide: Non-Uniform Sampling

- Standard approach to generating a random number for general p_i.
    1. Generate a uniform random number 'u' in the interval [0,1].
    2. Return the largest index 'i' such that u ≤ $\sum_{j=1}^{i} p_i$
- Conceptually, this divides interval [0,1] into non-equal-size pieces:



- Can sample from a generic discrete probability distribution in O(n).
- If you need to generate 'm' samples:
    - Cost is O(n + m log (n)) with binary search and storing cumulative sums.

Bonus Slide: Discussion of K-Means++

- Recall the objective function k-means tries to minimize:

$$\underbrace{f(\overset{\uparrow}{W}, \overset{\uparrow}{c})}_{} = \sum_{i=1}^{n} \|x_i - w_{c(i)}\|_2^2$$

all means   all assigments

- The initialization of 'W' and 'c' given by k-means++ satisfies:

$$\frac{E[f(W,c)]}{f(W^*, c^*)} = O(\log(k))$$

expectation over random samples

"Best" mean and clustering according to objective.

- Get good clustering with high probability by re-running.
- However, there is no guarantee that c^* is a good clustering.