Anomaly detection

Also known as "trova l'intruso".

Anomaly detection

Due to its practical use in the literature, we'll refer to anomalies also as *outliers*.

What is an outlier?

Outliers properties

Outliers are…

- **Inherently fuzzy**. An instance has a *degree* of outlierness, which we can threshold to decide whether an instance is an outlier or not.

Outliers properties

Outliers are...

- **Inherently fuzzy**. An instance has a *degree* of outlierness, which we can threshold to decide whether an instance is an outlier or not.
- **Data-dependent.** Outlier are exceptions to the data. But outliers themselves define the data...?

Outliers properties

Outliers are...

- **Inherently fuzzy**. An instance has a *degree* of outlierness, which we can threshold to decide whether an instance is an outlier or not.
- **Data-dependent.** Outlier are exceptions to the data. But outliers themselves define the data...?
- **Not noise.** Noise is *random*, outliers are *exceptional*.

Outliers properties

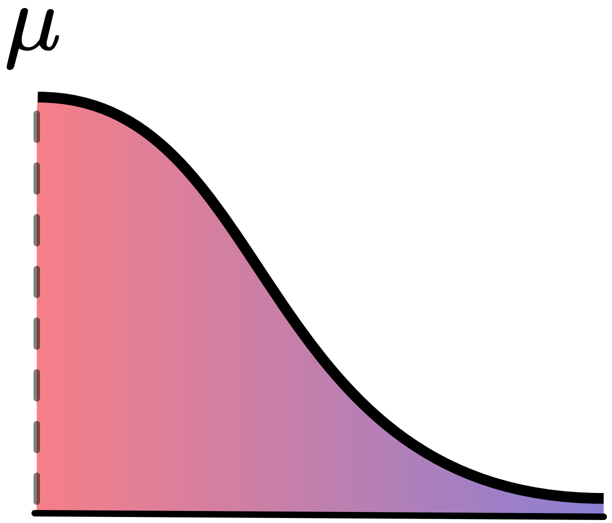
Outliers are…

- **Inherently fuzzy**. An instance has a *degree* of outlierness, which we can threshold to decide whether an instance is an outlier or not.
- **Data-dependent.** Outlier are exceptions to the data. But outliers themselves define the data…?
- **Not noise.** Noise is *random*, outliers are *exceptional*.
- **Mono- or multi-dimensional.** An outlier can be so on one just one dimension, or on multiple.

Defining outliers

Whatever the definition, we have two separate families of definitions:

- Something unusual. A penguin in this classroom.
- Something extreme. A cassata at a cake competition.



A Normal distribution $\mathcal{N}(\mu, \sigma)$.

Defining outliers

Whatever the definition, we have two separate families of definitions:

- Something unusual. A penguin in this classroom.
- Something extreme. A cassata at a cake competition.

Examples

We are given the census of Pisa.

- An outlier that is unusual?
- An outlier with extreme values?

Defining outliers

Whatever the definition, we have two separate families of definitions:

- Something unusual. A penguin in this classroom.
- Something extreme. A cassata at a cake competition.

Examples

- *Unusual*: a 95 y.o. Amazon native.
- *Something extreme*: a university professor.

The central problem with outliers

Outliers are, by nature, defined in terms of other instances. Whatever approach we use to detect them, we should take into account that they influence it as well.

The +1 problem. How many other "outliers" should I introduce in the data, before there are no more outliers?

Finding outliers: a 2-tier approach

Most algorithms use a two-tier approach:

1. **Grading** Define a grading function $\tilde{o}$ quantifying the *degree* of anomaly
2. **Thresholding** Define a thresholding function $\hat{o}$ to map the degree to a binary label

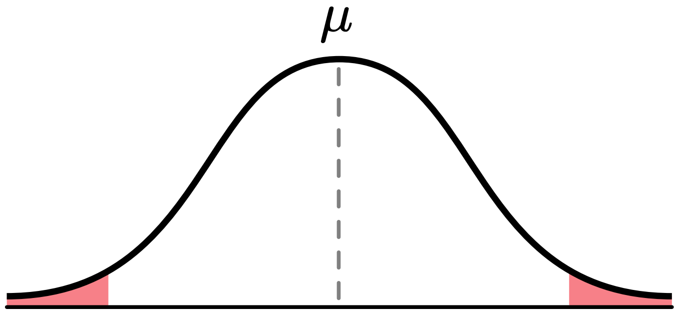
Axes of analysis

How to characterize outlier detection algorithms?

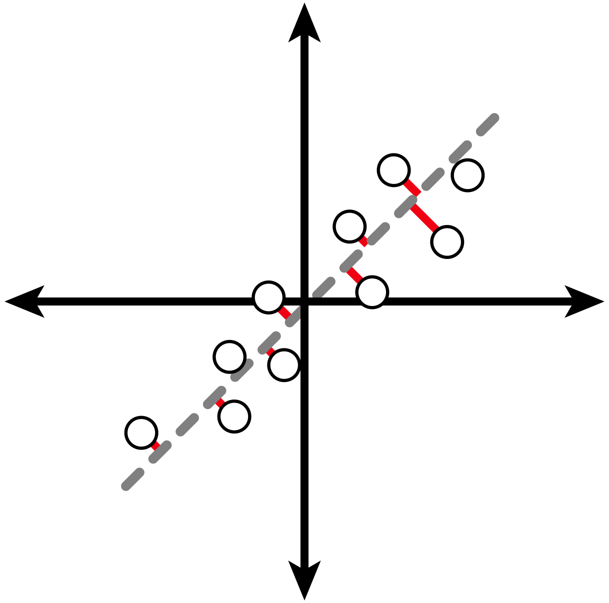
Axis	
Locality	Is the outlier *global* to the dataset, or *local* to a neighborhood?
Sensitivity	Is the algorithm heavily impacted by data with some particular characteristics?
Interpretability	Can we interpret why an instance is an outlier?

Defining unusual and extreme

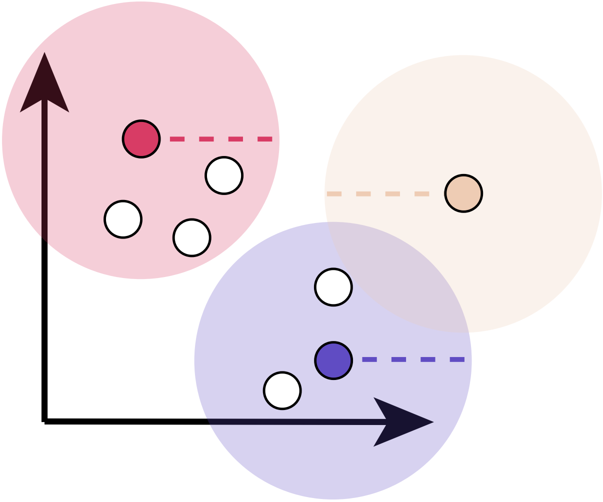
We define outliers by studying...



... the **distribution** of the data: $\tilde{o}$ is a function of the data distribution.



... the data **manifold**: $\tilde{o}$ is a function of the shape of the data.



... the **neighborhood**: $\tilde{o}$ is a function of the instance's neighbors.

Outliers and distributions

Data distributions offer a very natural and straightforward way of defining outliers, particularly when thinking of outliers as unusual occurrences.

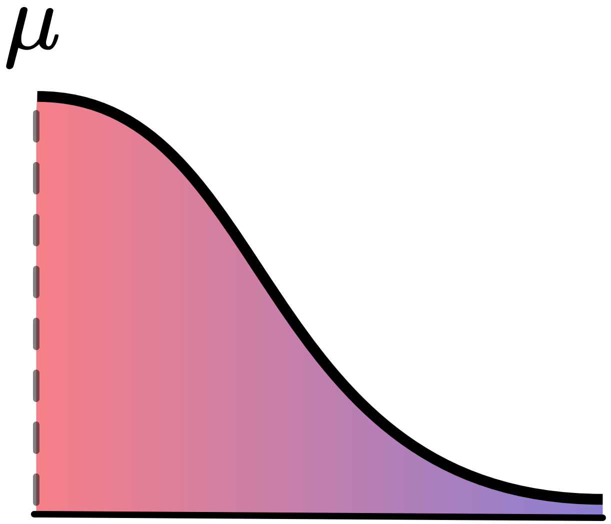
- ($\tilde{o}$) Scoring amounts to density estimation
- ($\hat{o}$) Thresholding amounts to critical value selection

Scoring, Normally: $z - scores$

What is the anomaly degree? The scaled distance from the mean.

Assumption: Data follows a normal distribution $\mathcal{N}(\mu, \sigma)$.

Idea: degree is given by weighted distance from the mean.

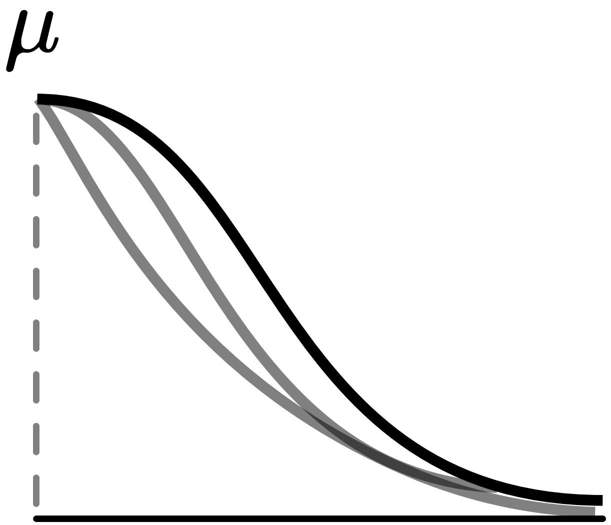


Degree of anomaly $\tilde{o}(x_i)$ of a sample x_i is $\tilde{o}(x_i) = \frac{x_i - \mu}{\sigma}$.

Strong relationship with t-SNE!

Tackling the +1 problem: Grubbs test

$z - scores$ generate sample-dependent outlier degrees $\tilde{o}(x_1), \ldots, \tilde{o}(x_n)$, but does not tackle the *+1 problem*. Grubb's test iterates over detected outliers, removing one layer of outliers at a time, until no more outliers are found.



A Normal $\mathcal{N}(\mu, \cdot)$ distribution at different bandwidths $\sigma_1, \ldots, \sigma_k$.

Designed for extremely small samples (<100), correction errors ensue when several iterations are performed.

Tackling the +1 problem: Grubbs test

Grubb's test iterates over detected outliers, removing one layer of outliers at a time, until no more outliers are found.

1. Find current outlier set $\hat{X}$
2. If $\hat{X} = \emptyset$, terminate
3. $X = X \setminus \hat{X}$, go to 1

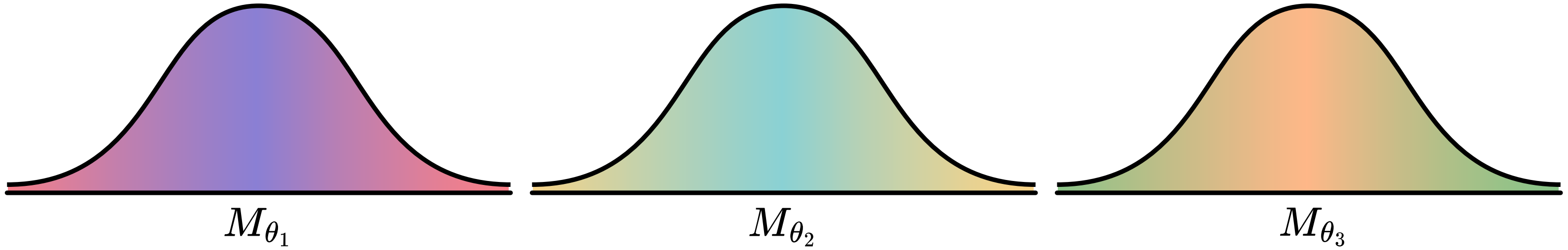
$z - scores$ and Grubbs test

Axis	
Locality	Global
Sensitivity	Outliers themselves influence the distribution, but can be removed (Grubbs)
Interpretability	Low: no reason other than "Not many similar instances"

Generalizing to distribution locality

Data may vary *locally*: subsets of the data each follow a different distribution.
Assumption: there exists a partition of the data, each block distributed according to a Normal distribution.

One of k models $M_{\theta_1}, \dots, M_{\theta_k}$ is sampled, each with a sampling probability m_i. Different distributions sample in different regions of the density, e.g., the data distribution may not be Normal, but some subspaces may.



A mixture of Normals $M_{\theta_1}, M_{\theta_2}, M_{\theta_3}$: each is sampled with probability m_1, m_2, m_3, respectively.

Mixture models

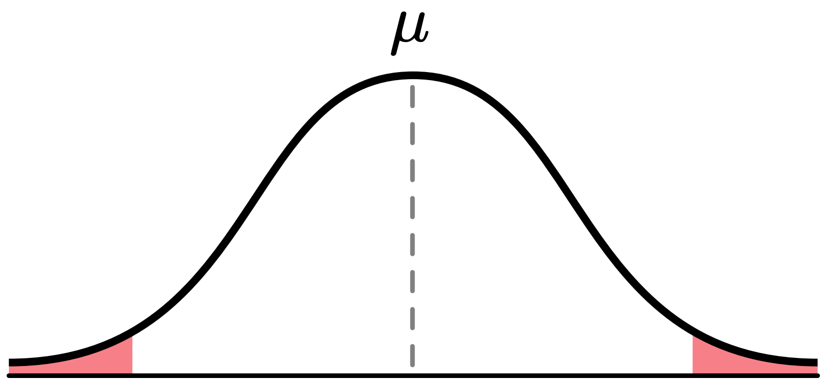
Axis	
Locality	Local
Sensitivity	Outliers themselves influence the distribution, can be unstable
Interpretability	Low: no reason other than "Not many similar instances"

Thresholding distributions

The critical values $\tilde{o}_x$ represent the density, i.e., relative likelihood of x: different thresholdings of $\tilde{o}_x$ yield different outliers. For some $\hat{o}, x$ is an outlier, for some others, it is not.

Choosing $\hat{o}$ is arbitrary, but some algorithms, such as Grubbs', define their own threshold

$$n \frac{\Sigma(x_i - \bar{X})^4}{(\Sigma(x_i - \bar{X})^2)^2}.$$



Tails of a Normal distribution.

Procedures for Detecting Outlying Observations in Samples, Frank E. Grubbs

Generalizing thresholding

$z - scores$ assume a Normal distribution, but often this is not the case. Yet, we can still identify *tails* of a distribution, and in turn, anomalies.

Markov inequality

For a variable X with positive values, and threshold β, it holds

$$\Pr[X > \beta] \leq \frac{\mathbb{E}[X]}{\beta}.$$

Thus, given an estimate of the variable's expected value, we can retrieve the inverse of an image of its cumulative distribution ($\Pr[X \leq \beta]$).

The tails of a distribution simply identify *extreme* events, thus only a category of anomalies applies.

Generalizing thresholding

$z - scores$ assume a Normal distribution, but often this is not the case. Yet, we can still identify *tails* of a distribution, and in turn, anomalies.

Chebychev inequality
For a variable X and threshold β, it holds

$$\Pr[| X - \mathbb{E}[X] |> \beta] \leq \frac{\sigma_X^2}{\beta^2}.$$

That is, the probability of deviation from the mean is inversely proportional to the deviation, and directly proportional to the variance.

The tails of a distribution simply identify *extreme* events, thus only a category of anomalies applies.

Modeling the data distribution

Assumption
The data follows a probabilistic process of the selected family.

Anomaly degree
Estimated density.

Thresholding
Critical value.

- Natural and straightforward definition of outliers
- Strong theoretical background
- Clear interpretation of the scores $\tilde{o}$, and clear definition of its thresholding function

- Sensitivity to outliers
- Sometimes unstable, especially in very high dimensions
- Limited expressivity
- Little interpretability of the result

Modeling the data manifold

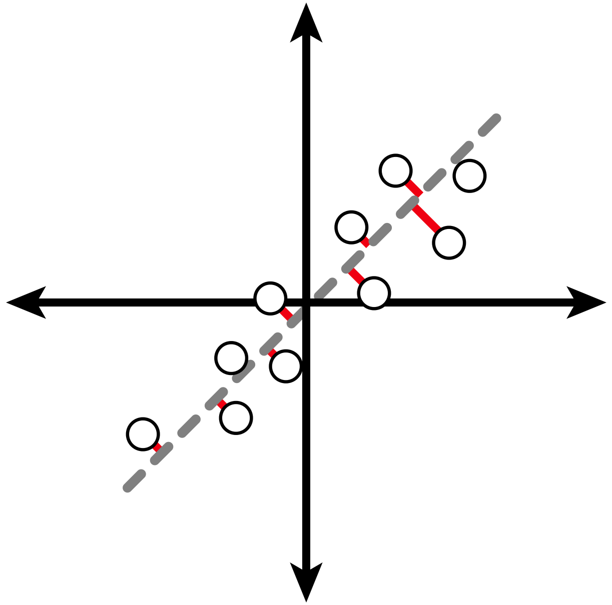
Distributional approaches define the density, but do not describe the data itself. $\tilde{o}$ is defined in terms of the manifold: does the given instance *lie* in the manifold? Just like the distributional approach, we must assume the manifold family.

To preserve the interpretability of our results, we stick to *linear* manifolds*.

*We won't.

Scoring in a manifold

By definition, the degree of anomaly an instance is its distance from the manifold.



A linear manifold (a plane), and distances of some instances to the manifold (in red).

Impossible manifolds and projections

A matrix A spans a linear space, thus every vector b in its spanned space is defined as a linear combination of A: $b = Ax$. For non fullrank matrices A, such a solution x may not exist. Thus, we need to *project* on the data manifold.

Least squares.

A least squares solution minimizes

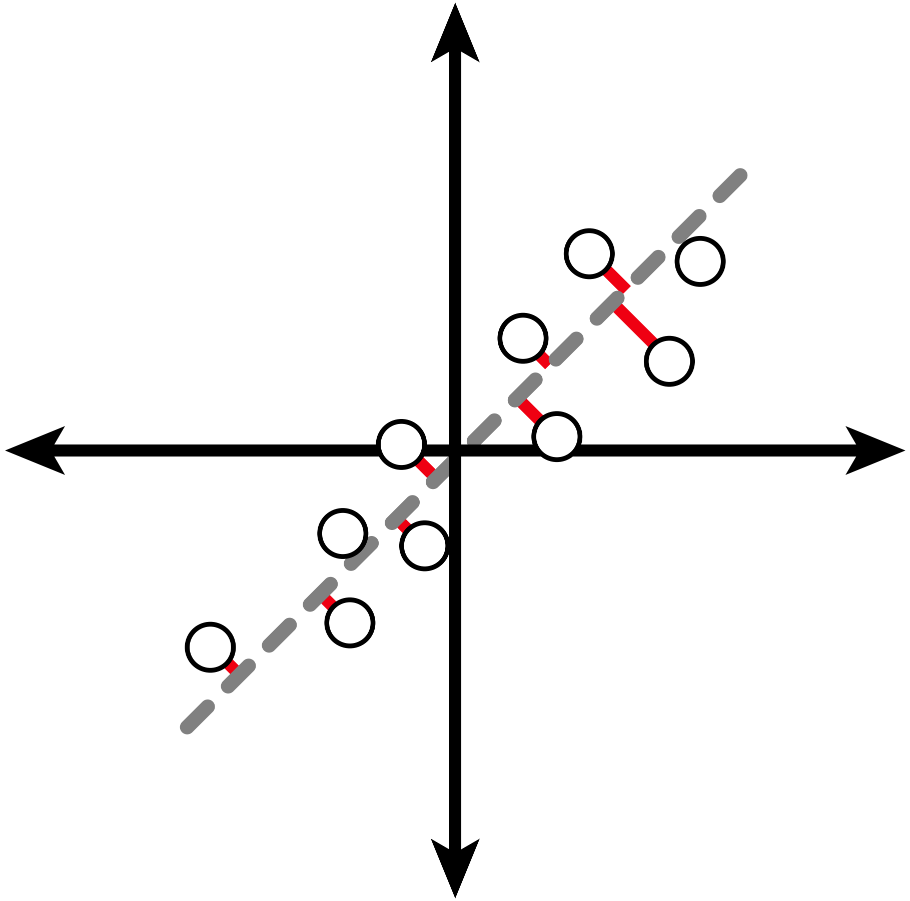
$$|| \, Ax - b \, ||_2^2.$$

Assumption: Least squares assumes a linear manifold, and squared norm as distance metric.

Grading in Least Squares

The least squares projection induces errors e_{x_i}, which can be used as outlier scores, i.e., $\tilde{o}(x_i) = e_{x_i}$.

Can we apply this to any dataset?

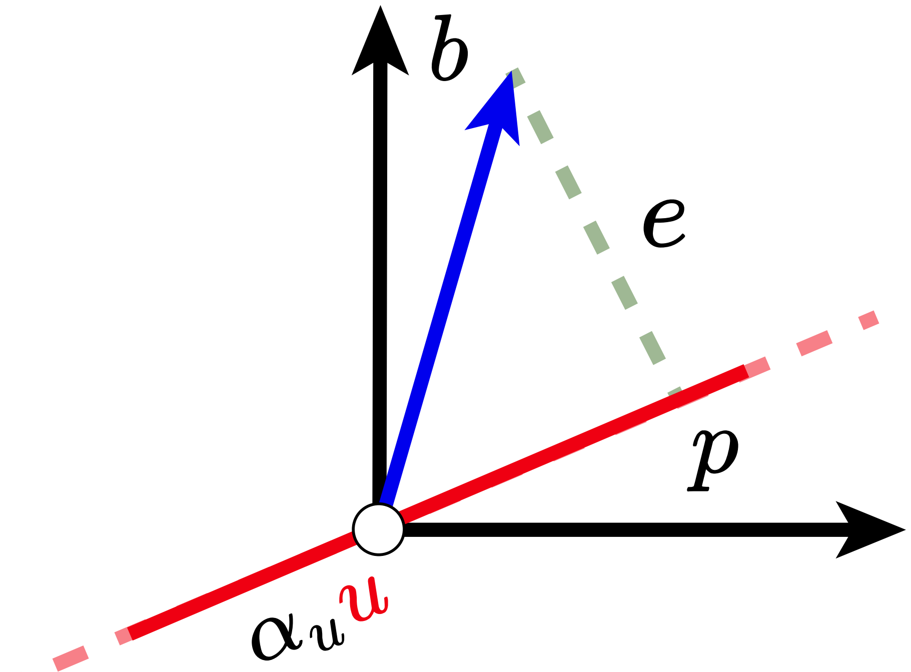


Scores $\tilde{o}(x_i)$ are given by the errors on the Least Squares approximation (in red in the picture).

Projections

The vector b is projected onto p, a vector on the linear space spun by A, and yielding an error vector $e = b - p$. The projection is orthogonal, thus it must hold $A^T e = 0$. Also, there exists a vector $\tilde{x}$ s.t. $A\tilde{x} = p$. Thus,

$$\begin{aligned} A^T e &= 0 \\ A^T(b - p) &= 0 \\ A^T(b - A\tilde{x}) &= 0 \\ A^T A\tilde{x} &= A^T b \\ \tilde{x} &= (A^T A)^{-1} A^T b \end{aligned}$$



Projection p of a vector b (in blue) on a subspace A: the error e is perpendicular to the subspace.

Least squares and collinearity

The formulation of the projection is thus

$$\tilde{x} = \overbrace{(\underbrace{A^T A}_{\text{sample covariance matrix}})^{-1} A}^{\textit{projection matrix } P}\, b,$$

which does not admit a unique solution for a singular $(A^T A)$, and is prone to instability for $A^T A$ nearly nonsingular. Since the sample covariance matrix $A^T A$ quantifies the collinearity of A, *least squares does not admit solutions for perfectly collinear data*. To make matters worse, when the data is nearly collinear, the computation is **unstable**.

Tackling collinearity: PCA

The instability of least squares is due to the data collinearity. A possible solution: decorrelate the data! Principal Component Analysis (PCA) does just this.

The cost: lower interpretability of the results.

Least Squares

Axis	
Locality	Global
Sensitivity	Strongly influuenced by outliers
Interpretability	Partial: which instances have lower degrees? What even is a "low" degree?

From descriptive to discriminative detection

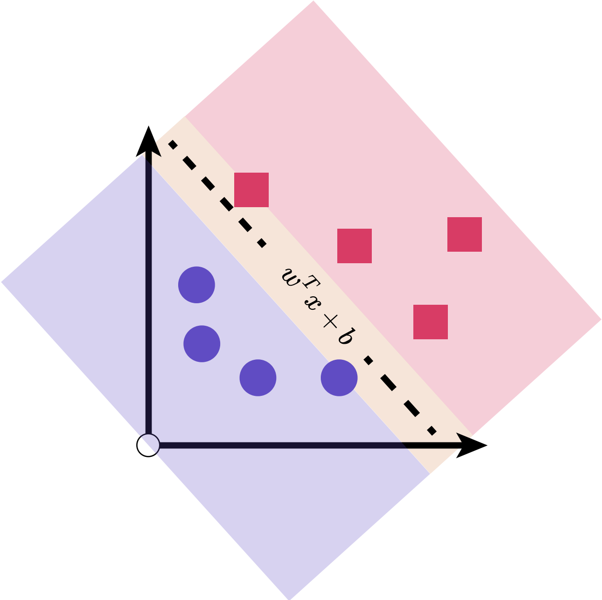
Manifold approaches *describe* the manifold by defining it in terms of its instances. What if instead, we were to define it in terms of...

- ... what sets it apart from other manifolds?
- ... what separates the manifold from other manifolds?
- ... what separates its instances from other instances?

(Linear) Discriminative outlier detection

Paradigm shift: we define the manifold as a *separating* manifold that separates the data from outliers.

- Assumption #1: I have some knowledge about which instances are outliers ($X^{\notin}$).
- Assumption #2: Outliers can be defined linearly with respect to the inliers ($X^{\in}$).



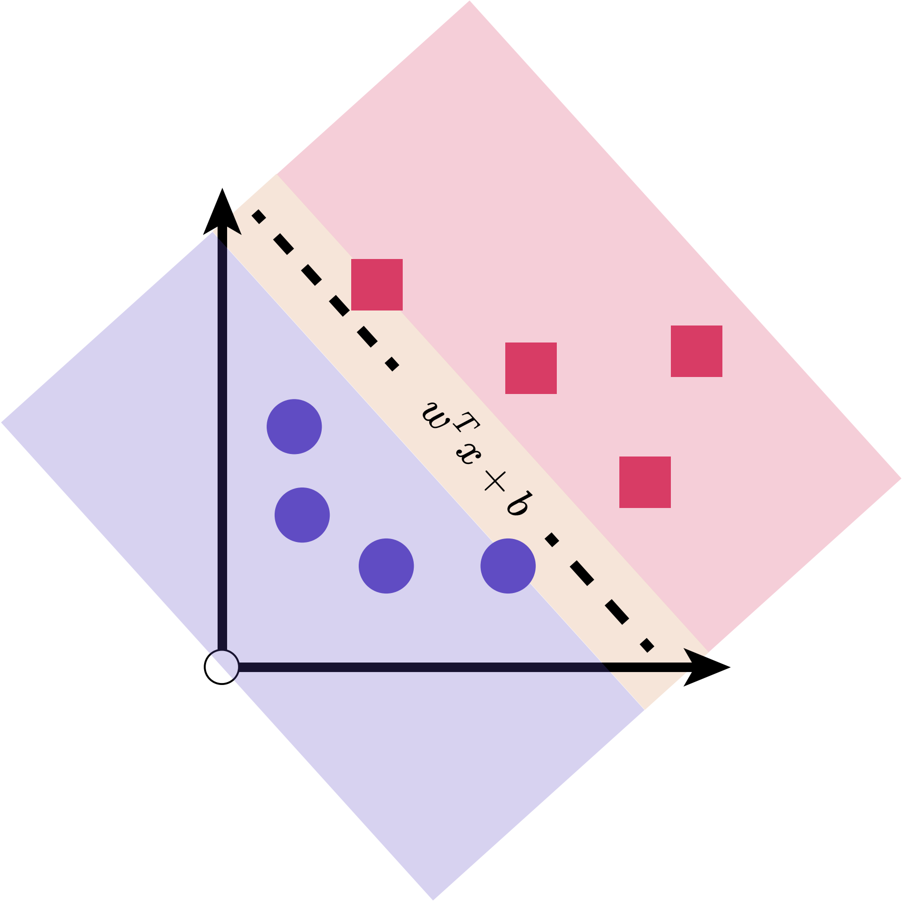
Inlier instances $X^{\in}$ (red squares) and a separating hyperplane $w^T x + b = 0$ separating them from outlier instances $X^{\notin}$ (blue circles).

(Linear) Discriminative outlier detection

Our goal: to best separate the outliers, that is, to **maximize** the distance between them and the inliers. In other words, to find a discriminative criterion maximizing the distance between inliers and outliers.

Two goals:

1. Find a formula for the *margin*
2. Maximize it



The *margin* (in beige) centered on the hyperplane separates inliers and outliers: we wish to maximize this!

Support Vector Machines

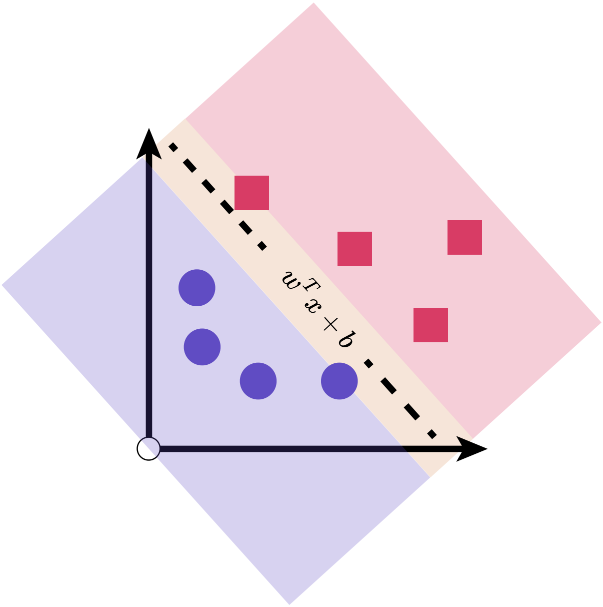
Let us define a hyperplane $w^T x + b = 0$ separating $X^{\in}$ and $X^{\notin}$, for which we have

$$\begin{cases} w^T x + b \geq +1 & \text{for } x \in X^{\in} \\ w^T x + b \leq -1 & \text{for } x \in X^{\notin} \end{cases}$$

Instances in the margin (called *support* instances/vectors) solve this for $w^T x + b = \pm 1$.

We can compact the two into

$$y^{\cdot}(w^T x + b) + 1 \geq 0$$



Instances and a separating hyperplane $w^T x + b = 0$. The two half-planes in red and blue are defined by $w^T x + b \geq +1$ and $w^T x + b \leq +1$, respectively.

Support Vector Machines

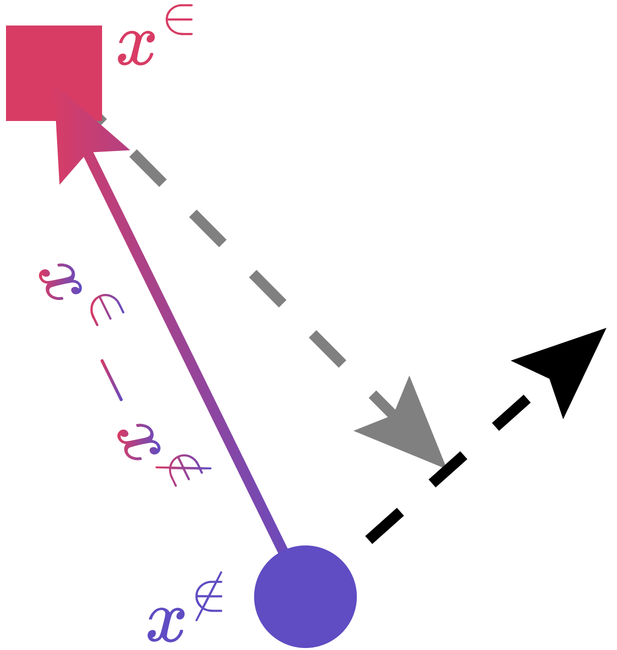
Geometrically, it is the projection of *margin* points onto a direction orthogonal to the margin:

$$(\hat{x}^{\in} - \hat{x}^{\notin}) \cdot \frac{w}{|| \, w \, ||},$$

which we can solve as

$$\frac{w \cdot \hat{x}^{\in} - w \cdot \hat{x}^{\notin}}{|| \, w \, ||} = \frac{(-b+1) - (-b-1)}{|| \, w \, ||} =$$

$$= \frac{2}{|| \, w \, ||}$$



Two instances $x^{\in}$ (red square), $x^{\notin}$ (blue circle), their difference $x^{\in} - x^{\notin}$ (in blue-to-red gradient), and a vector orthogonal to the margin (in black). The width of the margin is then the projection of the difference on such vector.

One-class Support Vector Machines

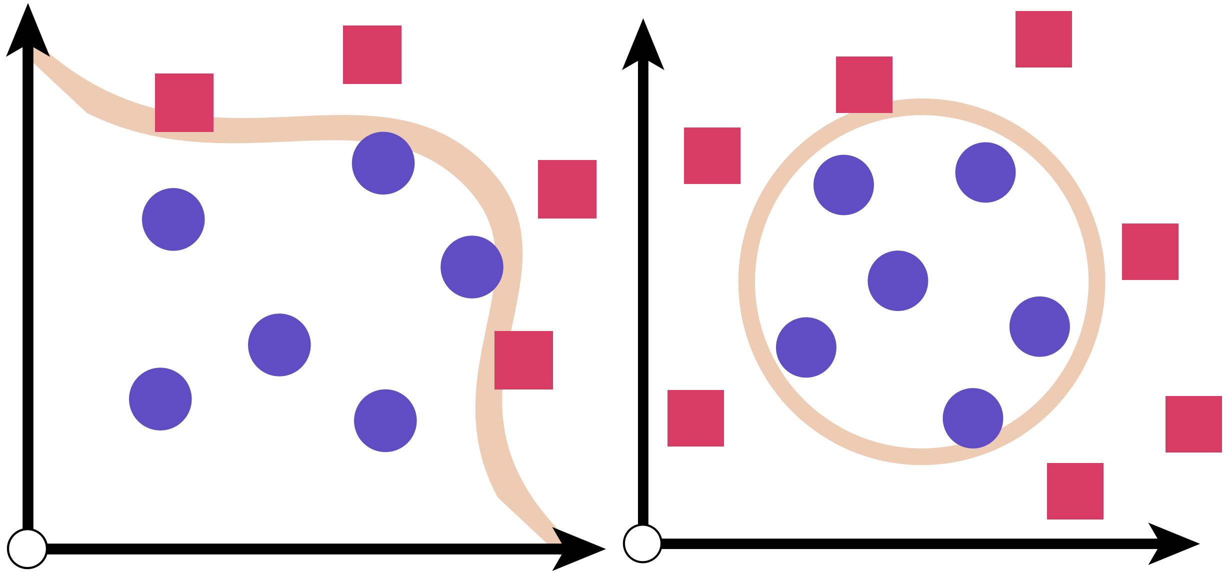
Solving analytically, we have that

1. The defining hyperplane w is a linear combination of the instances!
2. Some (hopefully many) instances have a zero coefficient λ_i, the others define (*support*) the hyperplane
3. The optimization takes the form $\Sigma_{i=1}^{n}\lambda_i - \frac{1}{2}\Sigma_{i=1}^{n}\Sigma_{j=1}^{n}\lambda_i\lambda_j y_i y_j \underbrace{x_i \cdot x_j}_{\textit{dot product!}}$

For a more in-depth derivation, see MIT OpencorseWare, 16. Learning: Support Vector Machines

Tackling linearity: the Kernel trick

Can we relax linearity without losing the interpretability of the algorithm? Yes, by changing the data itself, rather than the algorithm. We map the data from $\mathcal{X}$ to Φ, a space wherein instances are not strictly defined in terms of their features, but rather in terms of *inner products*, e.g., dot product, with other instances.



Kernel SVM: the margin can take nonlinear form.

What kernel Φ to choose?

There is a wide array of plug-and-play kernels we can use.

Kernel	Formulation	Description	Similarity
Linear	$x^T y$	Basic linear kernel	Angle-based
Radial basis	$exp(-\frac{\|\| x - y \|\|^2}{2\sigma^2})$	Exponentially decaying similarity	Distance-based
Polynomial	$(x^T y + c)^d$	Exponential kernel	Angle-based

Grading in One-Class Support Vector

Axis	
Locality	Global
Sensitivity	Choice of $X^{\notin}$: typically composed of negative instances
Interpretability	Yes! Support instances define the margin

Modeling the manifold

Assumption

The data lays on a (linear) manifold

Anomaly degree

Distance from the manifold

Thresholding

Unbounded and domain-dependent

- Flexible nonlinear manifold

- May be computationally unstable
- Strong manifold assumptions
- Possibly uninterpretable results

Modeling neighbors

Manifold-based algorithms are as flexible as the defined manifold. Like with mixture models, neighbor-based approaches reintroduce *locality*: outliers are defined in function of their neighbors:

- **Connectivity** An outlier is defined in terms of the connectivity to its neighbors
- **Concentration** An outlier is defined in terms of its neighbor concentration

Concentration is often defined as *density*: we use the former to remark that is is not a relative likelihood.

Modeling neighbors connectivity

Assumption: an instance is as much an inlier as it connected to other instances.

Each instance has a posting list of neighbors, from the closest to the farthest: the lower the aggregated position in other lists, the higher the connectivity degree.

- *Posting position* defines connectivity: it is **not** density
- Connectivity is asymmetric: I may be your closest instance, you may not be mine

$$\begin{bmatrix} 5 & 11 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 54 & 27 & \dots & 3 \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

Connectivity as a *postings* (not adjacency!) matrix A: $A_{i,j}$ is the $j-th$ nearest neighbor of instance i.

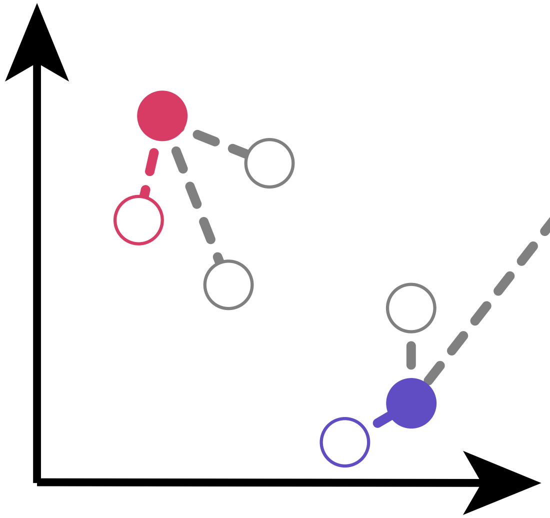
How do we populate the posting lists? $k-NN$!

Modeling neighbors connectivity

Assumption: an instance is as much an inlier as it connected to other instances.

Each instance has a posting list of neighbors, from the closest to the farthest: the lower the aggregated position in other lists, the higher the connectivity degree.

- *Posting position* defines connectivity: it is **not** density
- Connectivity is asymmetric: I may be your closest instance, you may not be mine



Neighbors at 1 of two instances (in red and blue): their neighbors circled of the same respective color.

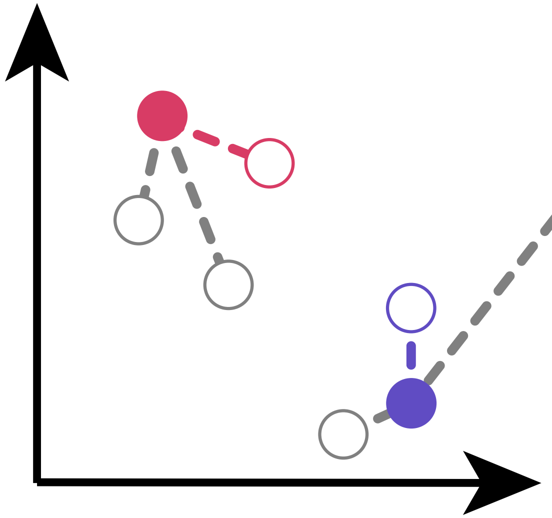
How do we populate the posting lists? $k - NN$!

Modeling neighbors connectivity

Assumption: an instance is as much an inlier as it connected to other instances.

Each instance has a posting list of neighbors, from the closest to the farthest: the lower the aggregated position in other lists, the higher the connectivity degree.

- *Posting position* defines connectivity: it is **not** density
- Connectivity is asymmetric: I may be your closest instance, you may not be mine



Neighbors at 2 of two instances (in red and blue): their neighbors circled of the same respective color.

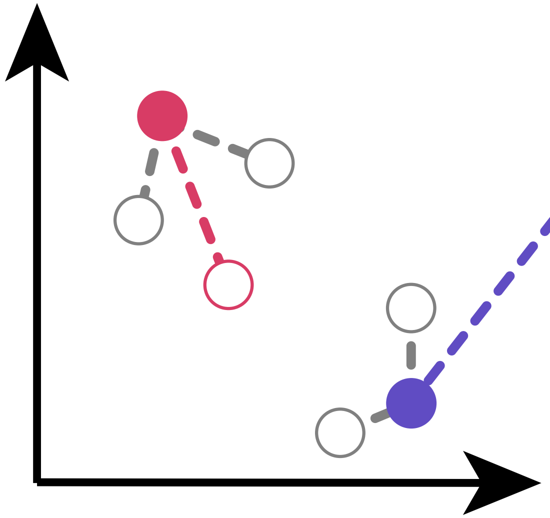
How do we populate the posting lists? $k-NN$!

Modeling neighbors connectivity

Assumption: an instance is as much an inlier as it connected to other instances.

Each instance has a posting list of neighbors, from the closest to the farthest: the lower the aggregated position in other lists, the higher the connectivity degree.

- *Posting position* defines connectivity: it is **not** density
- Connectivity is asymmetric: I may be your closest instance, you may not be mine



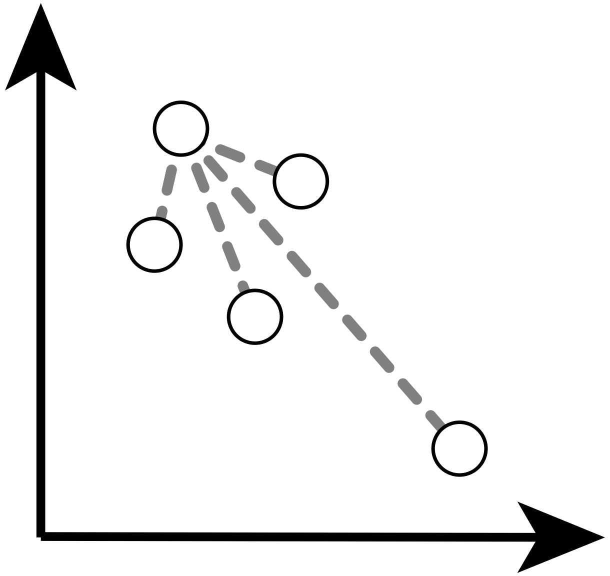
Neighbors at 3 of two instances (in red and blue): their neighbors circled of the same respective color.

How do we populate the posting lists? $k-NN$!

Grading neighbors connectivity

Posting matrices are often used as a base on which to measure different indices of connectivity, e.g.,

- *hub*: instance x_i is at least the $t-th$ neighbor of at least k other instances
- *popularity*: instance x_i is on average the $t-th$ neighbor of at least k other instances
- *ostracism*: instance x is, at worst, the $i-th$ neighbor of other k instances



An instance (top) and its connections to other instances in the dataset.

Thresholding neighbors connectivity

Connectivity lends itself to several thresholdings:

Position statistics

I threshold instances which are

- always
- on average
- never

the $(n - k) - th$ neighbor of other instances.

Neighbor statistics

I threshold instances which are at least the $i - th$ nearest neighbor of k instances.

Connectivity for hubs

Hub.
Instance x_i is at least the $t-th$ neighbor of at least k other instances.

Definition used by ODIN: given a posting matrix A, x_i is a hub if it appears at least k times in the first t columns of A. Hence, x_i is an outlier if the opposite is true:

$$\hat{o}(x_i) = \begin{cases} 1 & \text{if } \mid \{a \mid a \in A_{\neq i, \leq t}\} \mid < k \\ 0 & \text{otherwise.} \end{cases}$$

Outlier Detection using k-Nearest Neighbors Graph, Hautamaki et al.

Connectivity for popularity

Popular.
Instance x_i is, on average, the $t - th$ neighbor of other instances.

Given a posting matrix A, x_i is an outlier if, on average, is not less than the $t - th$ neighbor of other instances:

$$\hat{o}(x_i) = \frac{\sum_{i=0}^{n-1} \sum_{j=0, j \neq i}^{n-1} \mathbb{1}\{a_{i,j} = x_i\}}{n-1} > \alpha.$$

Posting matrices and connectivity

Posting matrices only allow us to appreciate connectivity as the *number* of connections, rather than their strength. If we were to superimpose a connectivity graph, this would only measure *how many* steps to take to connect to instance, and not *how long* should these steps be.

Grading neighbors concentration

Connectivity and concentration can be **approximated** through similar structures: we go from *postings* matrix to *distance* matrix!

To ease notation, we use a row-sorted distance matrix A_γ, so that row i holds increasing distances from instance x_i.

$$A = \begin{bmatrix} 0 & 2.28 & 0.16 & 0.21 \\ 2.21 & 0 & 1.21 & 3.91 \\ 0.16 & 1.21 & 0 & 0.76 \\ 0.21 & 3.91 & 0.76 & 0 \end{bmatrix}$$

$$A_\gamma = \begin{bmatrix} 0.16 & 0.21 & 2.28 \\ 1.21 & 2.28 & 3.91 \\ 0.16 & 0.76 & 1.21 \\ 0.21 & 0.76 & 3.91 \end{bmatrix}$$

A distance matrix A (top), and its row-sorted version A_γ (bottom). First column of 0s trimmed from A_γ.

Grading neighbors density: reach

An instance x has reach $\gamma^k(x)$ if the $k-th$ nearest neighbor is at distance γ^k, and average reach $\bar{\gamma}^k(x)$ if the average of $\{\gamma^1, \ldots, \gamma^k\}$ is $\bar{\gamma}^k(x)$.

Our row-sorted distance matrix A_γ is the *reach* matrix of the data! Indeed, A_γ defines both reach and average reach.

The average reach defines an empirical *approximate* concentration!

$$A_\gamma = \begin{bmatrix} \gamma^1(x_1) & \gamma^2(x_1) & \gamma^3(x_1) \\ \gamma^1(x_2) & \gamma^2(x_2) & \gamma^3(x_2) \\ \gamma^1(x_3) & \gamma^2(x_3) & \gamma^3(x_3) \end{bmatrix}$$

$$A_\gamma \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} \end{bmatrix} = \begin{bmatrix} \bar{\gamma}^1(x_1) & \bar{\gamma}^2(x_1) & \bar{\gamma}^3(x_1) \\ \bar{\gamma}^1(x_2) & \bar{\gamma}^2(x_2) & \bar{\gamma}^3(x_2) \\ \bar{\gamma}^1(x_3) & \bar{\gamma}^2(x_3) & \bar{\gamma}^3(x_3) \end{bmatrix}$$

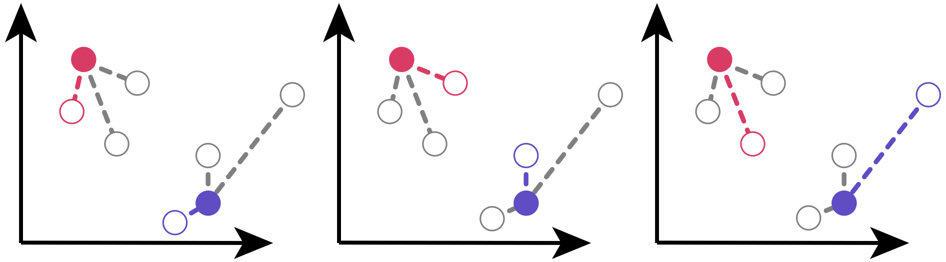
A_γ explicitly encodes reach (A_γ itself) and average reach.

Reach degrees: the reach ratio Factor

Assumption: Inliers have lower reach than their neighbors. We formalize this in a reach ratio

$$\tilde{o}^k_{i,j} = \frac{\bar{\gamma}^k(x_i)}{\bar{\gamma}^k(x_j)},$$

which is 1 for pairs x_i, x_j with equal k-neighbors concentration, and > 1 for instances with different concentrations, x_i laying in a sparser area of the space.

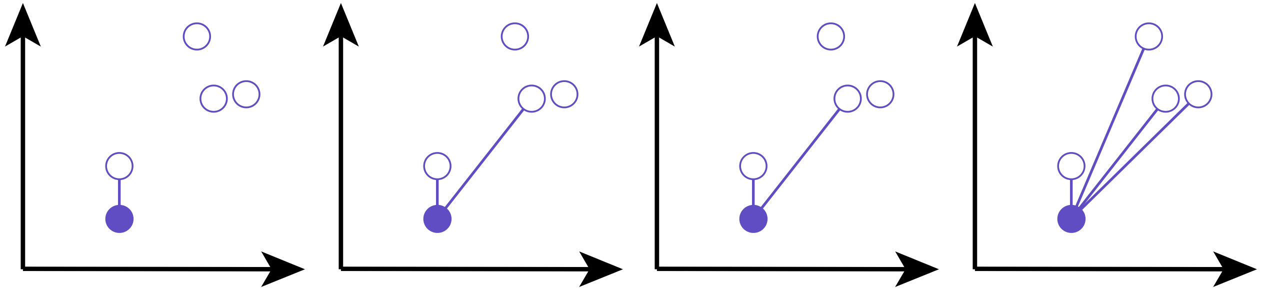


Reach at different ks: reach ratio factor averages ratios at different ks for pairs of instances x_i, x_j.

Reach degrees: Local Outlier Factor

Local outlier factor generalizes outlier factor by averaging the outlier factor over the neighbors of an instance:

$$\tilde{o}(x_i) = \Sigma_{x_j \in neigh(x_i)} \tilde{o}^k_{i,j}$$



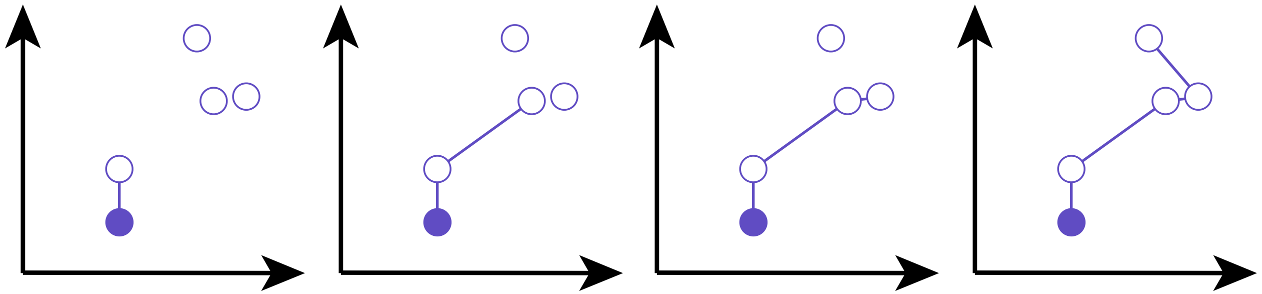
Neighbors at different k: Local Outlier Factor respects the posting matrix. It creates *clusters* of neighbors.

Reach degrees: Connectivity Outlier Factor

Connectivity outlier factor (COF) generalizes outlier factor by averaging the outlier factor over the *connected* neighbors of an instance:

$$\tilde{o}(x_i) = \Sigma_{x_j \in connect\_neigh(x_i)} \tilde{o}_{i,j}.$$

The connected neighbors of an instance x_i is recursively defined as the 1-nearest neighbor to the last element in the chain.

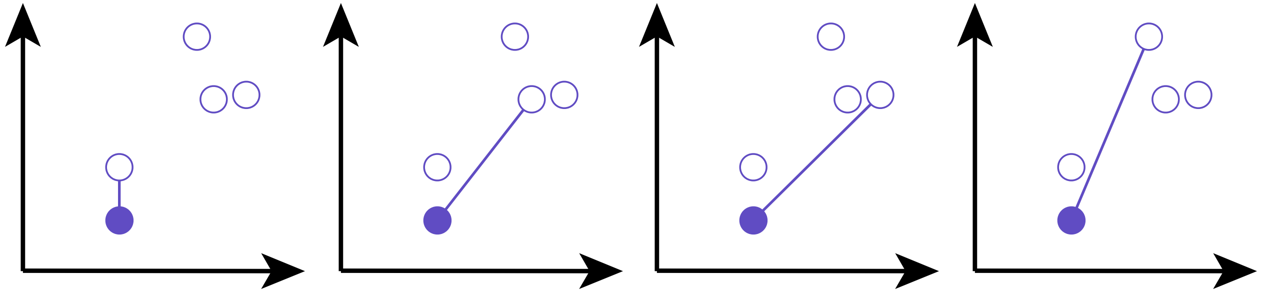


Neighbors at different k: Connectivity Outlier Factor does not respects the posting matrix. Rather, it creates *chains* of neighbors.

Reach degrees: $k - NN$ outlier factor

k-NN outlier factor (kOF) replaces the average reach at k ($\bar{\gamma}^k$) with the maximum reach at k ($\hat{\gamma}^k$):

$$\tilde{o}(x_i) = \gamma^k(x_i).$$



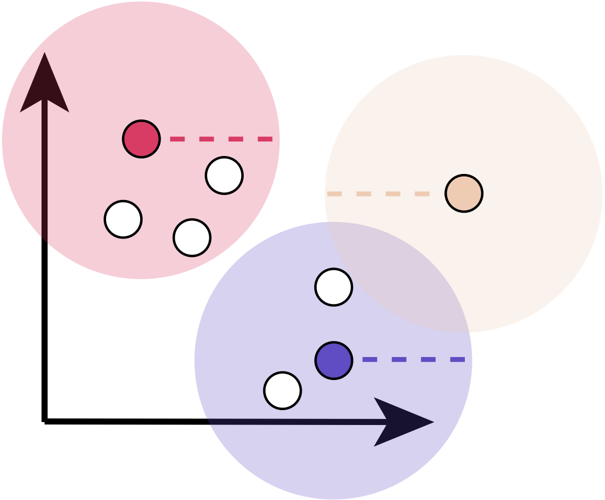
Neighbors considered at different k.

Efficient algorithms for mining outliers from large data sets, Ramaswamy et al.

Degrees of neighbors concentration

Reach degrees approximate space concentration with (inverse) reach. Rather than pick a k, we can swap in a more natural definition of concentration: instances found per unit of space. Even better, instances found within an hypersphere $B(\cdot, \varepsilon)$ of a given radius ε, and centered around $\cdot$.

Assumption: outliers have lower concentration than their neighbors.



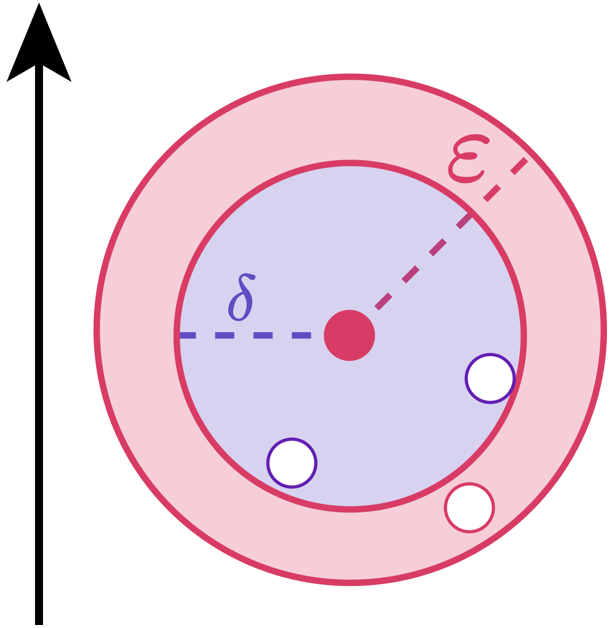
Instances, and some ε-hyperspheres centered on them.

B stands for ball, also often referred to as ε-ball.

Degrees of two-radii concentration

We compute concentration on a two-radii approach:

- **concentration radius** ε: determines the hyperspheres $B(x_i, \varepsilon)$ estimating concentration $c^{\varepsilon}(x_i)$ of x_i within a radius ε
- **neighborhood radius** δ: proportional to ε, determines the neighborhood B_i of x_i as the instances laying within $B(x_i, \delta)$



The two radii ε, δ: the former is used to estimate *concentration*, the latter to choose which neighbors to compare concentration against. **Note: δ may also be larger than ε!**

Degrees of two-radii concentration

Like reach-based concentration, degrees are defined on a basis of comparisons between some degree of an instance, and its neighbors:

$$\tilde{o}(x_i) = \bar{c}^{\varepsilon}(B_i) - c^{\varepsilon}(x_i),\ \text{with } \bar{c}^{\varepsilon}(B_i) = \frac{\Sigma_{x_j \in B_i} c^{\varepsilon}(x_j)}{\mid B_i \mid}$$

that is, two-radii concentration compares the concentration of an instance, with the concentration of its neighbors. For

- $\tilde{o}(x_i) >> 1$ neighbors have a much higher concentration
- $\tilde{o}(x_i) \rightarrow 0^+$ neighbors have a much lower concentration

LOCI: Fast Outlier Detection Using the Local Correction Integral, Papadimitriou et al.

Thresholding connectivity factors

Unlike distributional approaches, connectivity factors rely on arbitrary densities and distances, both of which are domain dependent and of unclear interpretation.

Grading connectivity factors

Axis	
Locality	Local
Sensitivity	Choice of neighborhood, connectivity parameter
Interpretability	Partial: can inspect what instances lead to different reaches

Fast neighborhood estimation

Neighbor approaches rely on **expensive** neighborhood functions, e.g., k-NN, and in turn build anomaly degrees on the basis of different assumptions on said neighborhoods: the neighborhoods determine the anomaly degree *post-hoc* through different cheap scoring functions:

- LOF: ratio of neighborhood reach
- COF: ratio of chain reach
- LOCI: difference of concentrations

Idea: what if instead, we build **simpler** and **faster** neighborhoods?

Fast neighborhood estimation

Wisdom of the crowd

Even if approximated, if I sample enough neighborhoods of variable quality, on aggregate I can achieve a representative neighborhood.

Outlier degree

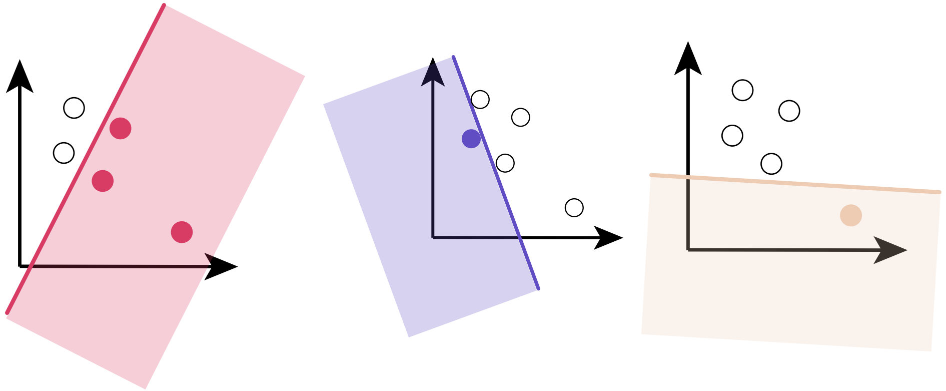
If neighborhood definitions induce an outlier degree, then we can estimate the outlier degree directly from the neighborhoods.

Isolation Forest, Fei Tony Liu et al.

Neighborhoods as hyperplanes

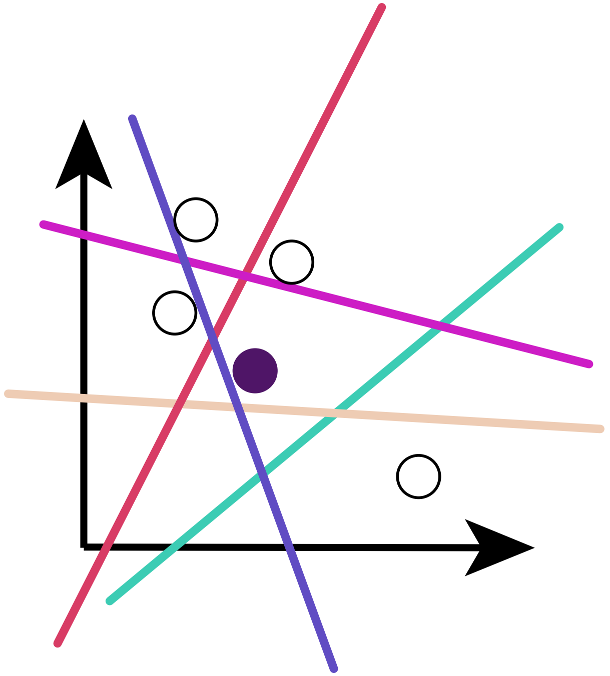
Wisdom of the crowd

Random sampling on a distribution of hyperplanes.



Outlier degree

The number of hyperplanes needed to define the neighborhood.



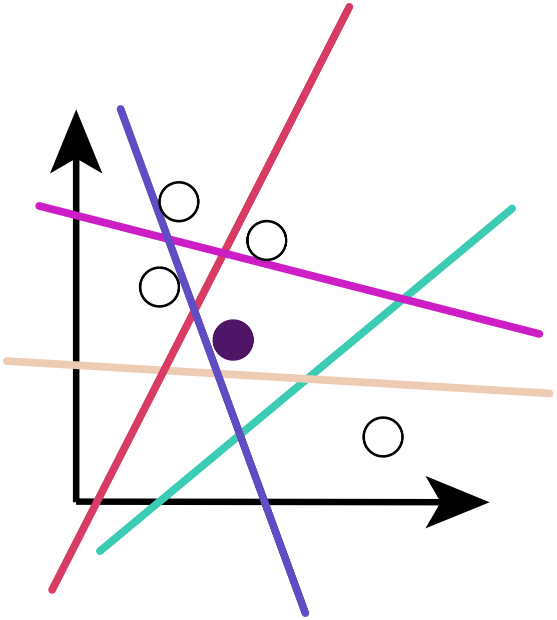
Isolation Forest, Fei Tony Liu et al.

Isolation tree

An *isolation* tree t is a random tree which randomly partitions the space into a set of blocks.

- Splits are sampled randomly
- Tree grows up to a predefined height, or until all leaves contain one instance

Outlier degree $\tilde{o}^t(x_i) = \dfrac{path(x_i, t)}{c}$.



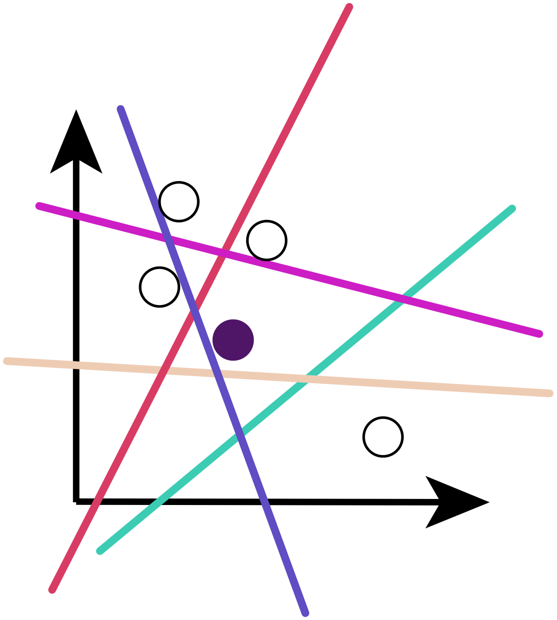
Isolation Forest, Fei Tony Liu et al.

Isolation forest

An isolation forest T is comprised of several isolation trees, further sampling the hyperplane space.

Outlier degree

$$\tilde{o}(x_i) = 2^{-\frac{\sum_{t \in T} path(x_i, t)}{|T| c}}$$



Isolation Forest, Fei Tony Liu et al.

Grading with isolation forests

Axis	
Locality	Global and local
Sensitivity	Dataset noise can be interpreted as outlier
Interpretability	Yes! Splits induced by the tree, if the tree is univariate

Modeling connections

Assumption
Outliers have a peculiar neighborhood

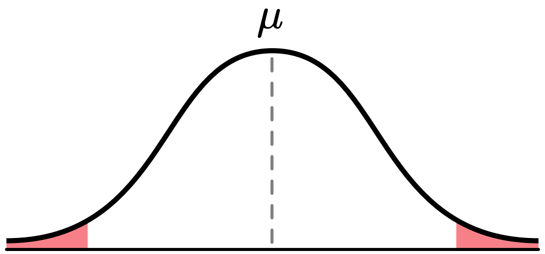
Anomaly degree
Distance from neighborhood

Thresholding
Distance from neighborhood

- Extremely flexible

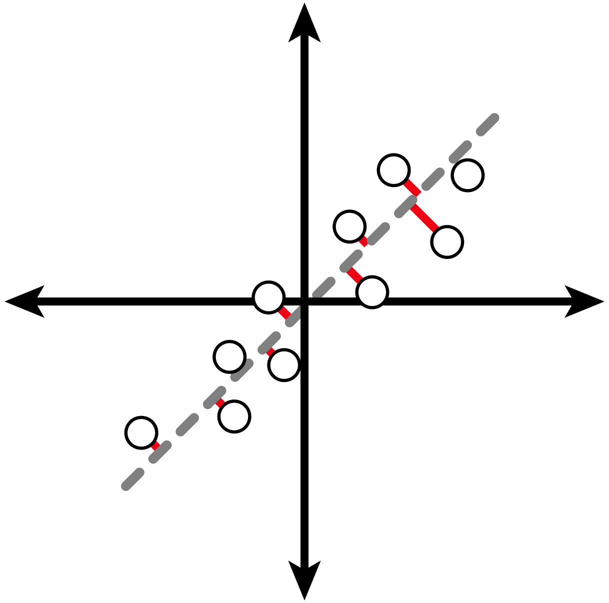
- Sensitive to hyperparameters
- Need to define a proper distance function

Finding outliers



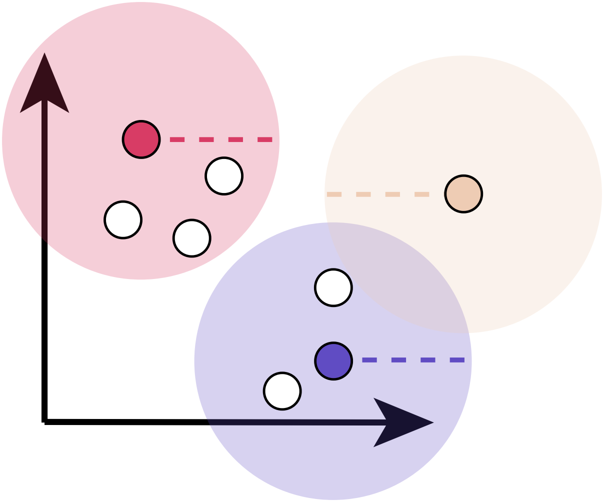
Assumption: The data distribution.

Thresholding: Critical value.



Assumption: The manifold family.

Thresholding: Distance to the manifold.



Assumption: Distances define anomalies.

Thresholding: Connection to neighbors.

Finding outliers

Data distribution

- Natural and straightforward definition
- Strong theoretical background
- Clear interpretation of the scores $\tilde{o}$

- Sensitivity to outliers
- Sometimes unstable
- Limited expressivity
- Little interpretability of the result

Data manifold

- Flexible and powerful

- May be computationally unstable
- Strong manifold assumptions
- Possibly uninterpretable results

Data neighbors

- Extremely flexible

- Sensitive to hyperparameters
- Need to define a proper distance function

Going meta: cluster approaches

We have studied clustering as a task aimed at discovering groups, which we can, in turn, leverage to discover outliers!

- Distributional approaches on separate clusters
- Reach approaches based on clustering, rather than single instances
- Connectivity approaches w.r.t. cluster centers, rather than other instances

References

Anomaly Detection, Charu C. Aggarwal. Second edition.

Topic	Sections
Anomaly detection.	1.3.1-4
Distributional approaches.	2.2, 2.4.1, 2.5
Manifold approaches.	3.2, 3.3
Connectivity approaches.	4.2, 4.3, 4.4, 4.5