Robust Artificial Intelligence

Reading Seminar; Tsinghua University

Thomas G. Dietterich, Oregon State University

tgd@cs.orst.edu

The Class So Far

- Lecture 1: Calibrated Probabilities (Closed World)
- Lecture 2: Thresholding Confidence Indicators (Closed World)
- Lecture 3: Open World

Lecture 3: Open Category Detection

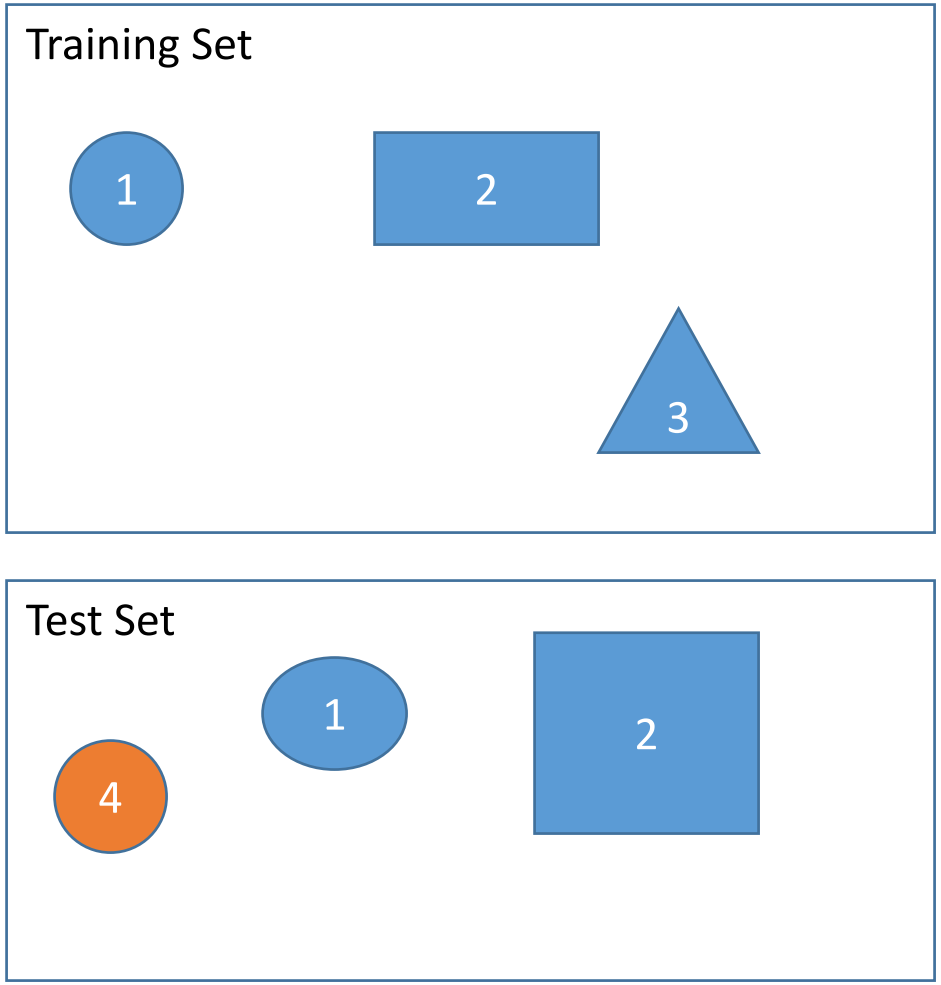
- Training:
  - Data: $(x_1, y_1), \dots, (x_N, y_N)$ drawn from D_0
  - $y_i \in \{1, \dots, K\}$
- Testing:
  - Data: Mixture D_m of data from D_0 and D_a
  - $(x, y) \sim D_a$ belong to new classes not seen during training ("alien categories")
- Goal:
  - Given a query x_q, does it belong to D_a or D_0?
  - If from D_a, REJECT as alien
  - Else classify using a classifier trained on D_0 data

Papers for Today

- Bendale, A., & Boult, T. (2016). Towards Open Set Deep Networks. In CVPR 2016 (pp. 1563–1572). http://doi.org/10.1109/CVPR.2016.173
- Liu, S., Garrepalli, R., Dietterich, T. G., Fern, A., & Hendrycks, D. (2018). Open Category Detection with PAC Guarantees. *Proceedings of the 35th International Conference on Machine Learning, PMLR, 80,* 3169–3178. http://proceedings.mlr.press/v80/liu18e.html
- Shafaei, A., Schmidt, M., & Little, J. (2018). Does Your Model Know the Digit 6 Is Not a Cat? A Less Biased Evaluation of “Outlier” Detectors. arXiv 1809.04729

Challenges of Open Category Recognition

- Discriminative training seeks the minimum information sufficient to separate class k from the other classes $\{1, \ldots, k-1, k+1, \ldots, K\}$
- Feature selection based on discriminative power (e.g., mutual information) may discard features that would be important for detecting aliens

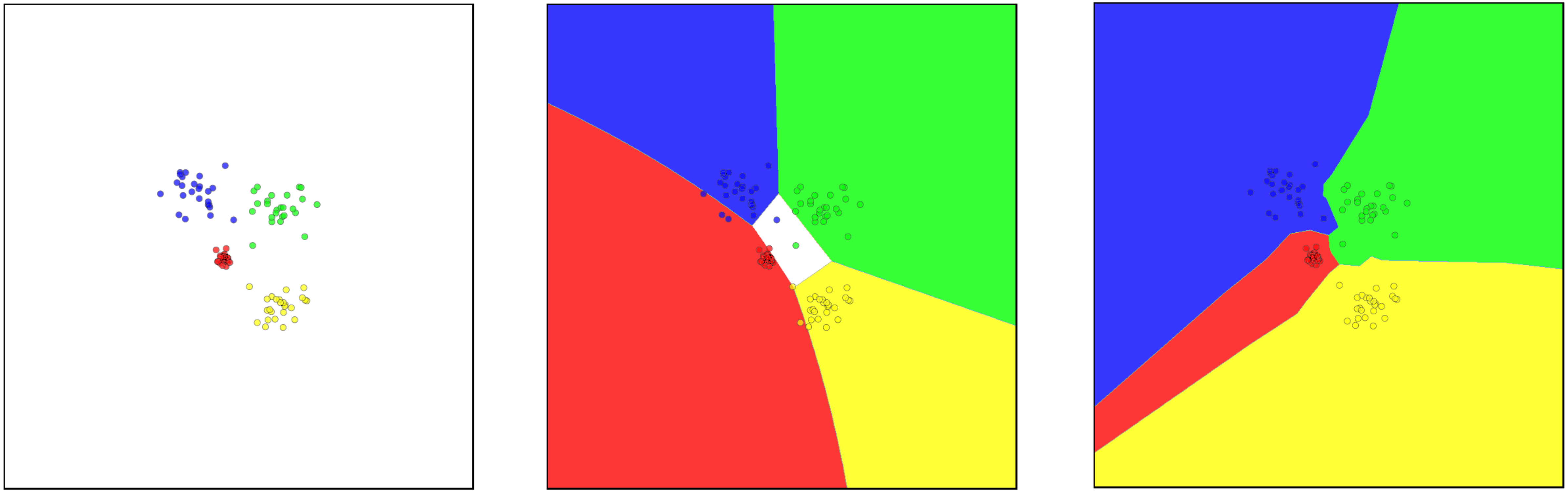


Discarding Useful Features

- In my insect identification project, we converted images to monochrome because experiments showed that color was not needed for accurate classification
- Claim: It is never safe to discard features when looking for anomalies/novelty

Visualizing the Challenges

- 4-Gaussian Data Set



SVM(MCBIN) NN

Mendes-Junior et al.

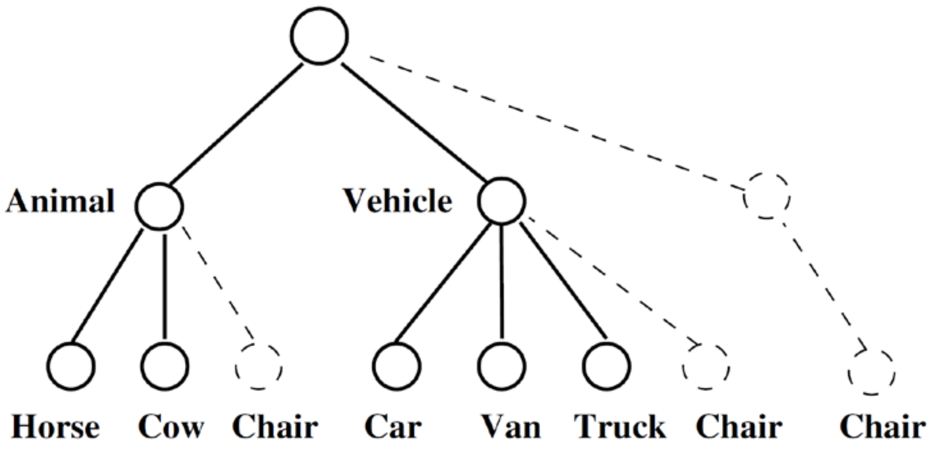
Ideal Method

- Estimate the density of each known class $P(x|k)$
- If $\max_k P(x_q|k) > \tau$ then classify as k
- Else REJECT

- This could be further improved if we had a theory of the classes
  - Typical separation from one another
  - Distinctiveness (how well can they be discriminated from each other)
  - Component parts (e.g., new class of vehicle will probably have wheels)

Approximating the Ideal

- Salakudtinov et al. (2011)
  - Hierarchical probabilistic clustering model
  - Each node contains a model in terms of subparts
  - Each subpart has an appearance model in terms of low-level filters learned from 1 million+ web images
- Classification:
  - Deciding where to put x_q
  - As an instance of an existing concept
  - As an instance of a new concept
- Not evaluated as an open category model



Multi-layer RBM

Salakhutdinov et al, 2011

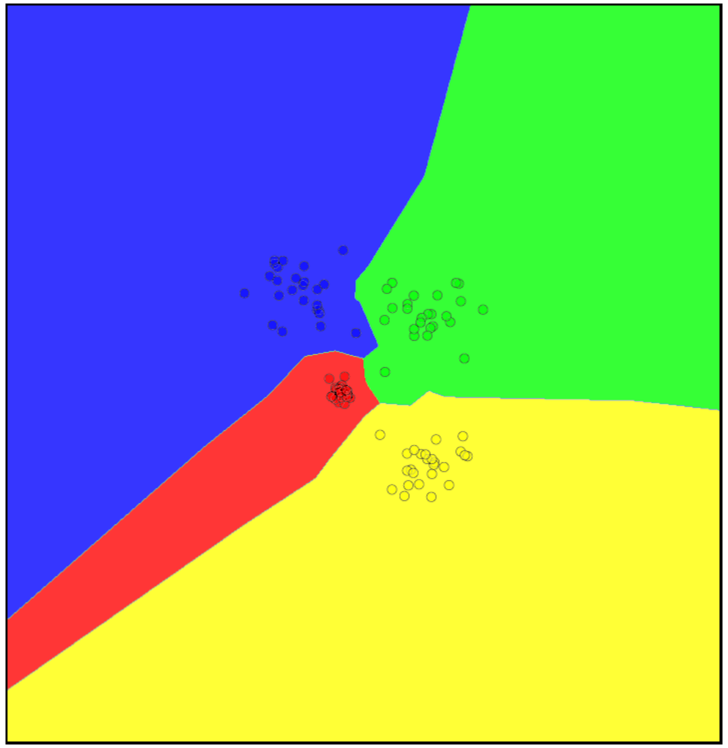
General Purpose Approaches

- Thresholding Standard Classifiers
- Anomaly Detection Filter
- Supervised Learning with Synthetic Open Space Examples
- “Other”

Method 1: Thresholding Standard Classifiers

- Let $f(x) = [\hat{p}(y = 1|x), \ldots, \hat{p}(y = K|x)]$
- Let $\hat{p}_{max} = \max_k \hat{p}(y = k|x)$
- REJECT if $\hat{p}_{max} < \tau$
- Else predict $\arg\max_k \hat{p}(y = k|x)$

- This does not work well because it focuses on the areas near the decision boundaries



Mendes-Junior et al.

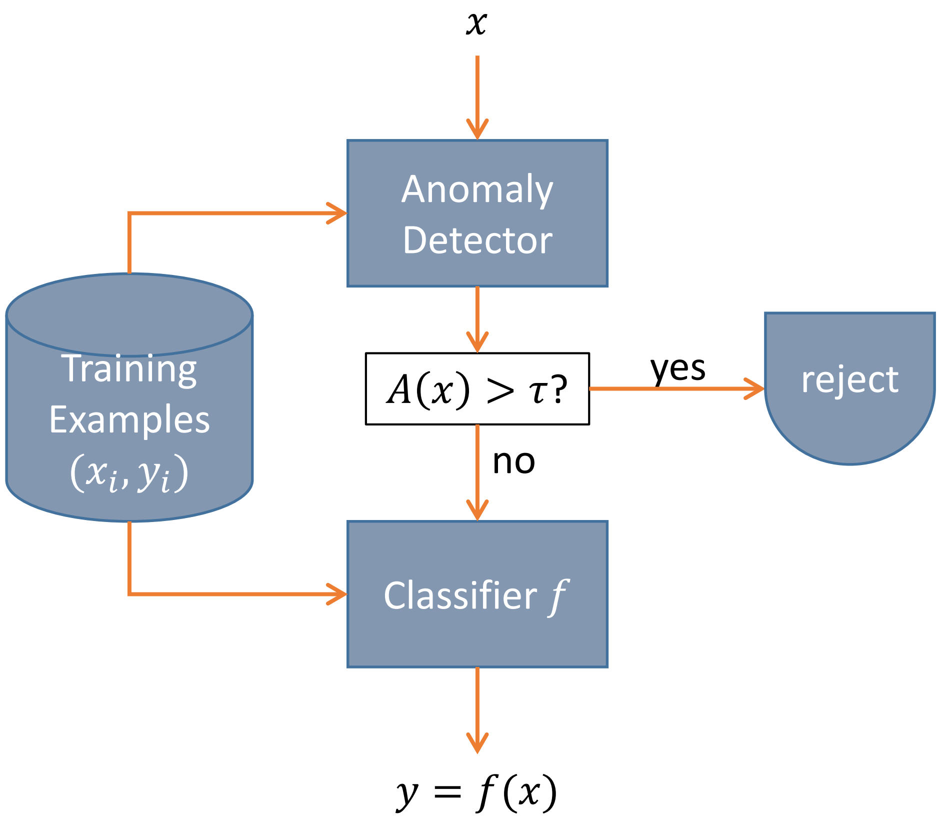
Method 1’: Thresholding 1-vs-rest Binary Classifiers

- For each training class k, learn a binary classifier $f_k(x) = P(y = k|x)$ versus $y \neq k$
- Set a threshold τ_k for each class
- If $f_k(x) < \tau_k$ for all k, then REJECT
- Else classify $\hat{y} = \arg\max_k f_k(x)$

- This also doesn’t work well, because it focuses on the one-vs-rest decision boundaries
- But by setting τ_k large enough, it works better than thresholding the multinomial logit (softmax)

Method 2: Series Anomaly Detector

- Per-Class Anomaly Detectors:
  - Distance Based AD (distance to nearest neighbor)
  - One-Class SVM
  - Extreme Value Distribution Models
- Multiclass Anomaly Detectors:
  - Kernel Null Space Method
  - Neural Fisher Discriminant



Extreme Value Distribution Anomaly Detection

- The Weibull Distribution is one possible model of the sampling distribution of the max
  - Repeat for $i = 1, \ldots, N$
    - Draw sample S_i of size n from distribution D
    - Let $x_i = \max_{x \in S_i} x$
  - The points $\{x_1, \ldots, x_N\}$ exhibit an extreme value distribution
- The CDF of the Weibull is $F(x) = 1 - \exp\left(\frac{\|x-\tau\|}{\lambda}\right)^{\kappa}$
  - τ "location parameter"
  - λ "scale parameter"
  - κ "shape parameter" $\kappa \in [1,2]$

Extreme Value Distribution Anomaly Detection

- Bendale & Boult:
  - Let μ_k be the mean of the data points in class k
  - Let $\{x_1, \dots, x_N\}$ be the N points in class k most distant from μ_k
  - Fit a Weibull distribution to them
- The probability that x_q is an alien with respect to class k is

$$F(\|x_q - \mu_k\|)$$

  - This is heuristic
  - We could have just set a threshold on $\|x_q - \mu\|$ but this attempts to calibrate the tails of the distribution for each class so they are all on the same scale
- Let $P_a(x_q) = \min_k F_k(\|x_q - \mu_k\|)$
- If $P_a(x_q) > \tau$ then REJECT

OpenMax (Bendale & Boult, 2015)

- Let $\ell_1, \ldots, \ell_K$ be the activations of the penultimate layer (the input to the softmax)
- Sort in descending order and index using k
- $\ell_0 \coloneqq 0$
- For $k = 1, \ldots, C$
  - Let $\omega_k = 1 - \frac{C-k}{k} \exp\left(\frac{\|x_q - \tau_k\|}{\lambda_k}\right)^{\kappa_k}$
  - $\ell_0 \coloneqq \ell_0 + (1 - \omega_k)\ell_k$
  - $\ell_k \coloneqq \omega_k \ell_k$
- Output $\mathrm{Softmax}(\ell_0, \ell_1, \ldots, \ell_K)$
- If class 0 has highest probability, then REJECT

Kernel Null Space Method

(Bodesheim, Freytag, Rodner, Kemmler, Denzler, CVPR 2013)

- Let N_k be the number of training examples for class k
- Assume $N_k < d$ (the input dimension)
- Assume the training examples are linearly independent
- Then there exists a linear transformation that maps all examples in class k to a unique point t_k
- Use $\min_{\mathrm{k}} \|x_q - t_k\|$ as the anomaly score

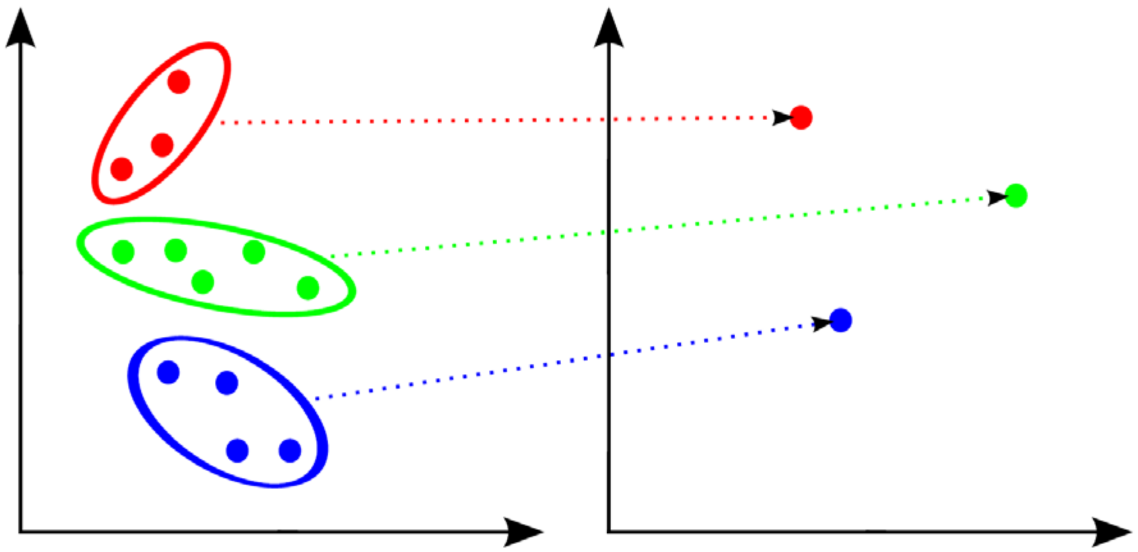


Figure 2. Visualization of NFST using three classes mapped from the input space (*left*) to the null space (*right*), adapted from [7].

Kernel Null Space Method (2)

- Let $k(x_i, x_j)$ be a kernel function whose feature mapping is $\phi(x)$
- If we use a high-dimensional kernel (e.g., the Gaussian) then $N_k \ll d_\phi$, so we can always compute this null space mapping

- Local version: Compute the null space mapping using only the M nearest neighbors to x_q (where, e.g., $M = 750$)
- Question: What does the null space mapping do to the empty space?

Neural Fisher Discriminant (Hassan & Chan, arXiv 2018)

- Learn an encoding network g such that
  - $\mu_k = \frac{1}{N_k} \sum_{i=1}^{N_k} g\left(x_{i,k}\right)$ "mean latent space value"
  - $\sum_{k=1}^{K} \sum_{i=1}^{N_k} \left\| g\left(x_{i,k}\right) - \mu_k \right\|^2$ "intra-class spread" is minimized
  - $\min_{k,k'} \left\| \mu_k - \mu_{k'} \right\|^2$ "between class spread" is maximized
  - They train using minibatches to compute the above
- Compute the anomaly score for x_q as

$$A\left(x_q\right) = \min_k \left\| g\left(x_q\right) - \mu_k \right\|$$

Method 3: Supervised Learning with Synthetic Examples

- Train a GAN and use it to generate synthetic alien data points in the open space
- Train a multiclass classifier to discriminate the K known classes from these synthetic examples
- Classify into the most likely class

Ge, Demyanov, Chen & Garnavi: Generative OpenMax

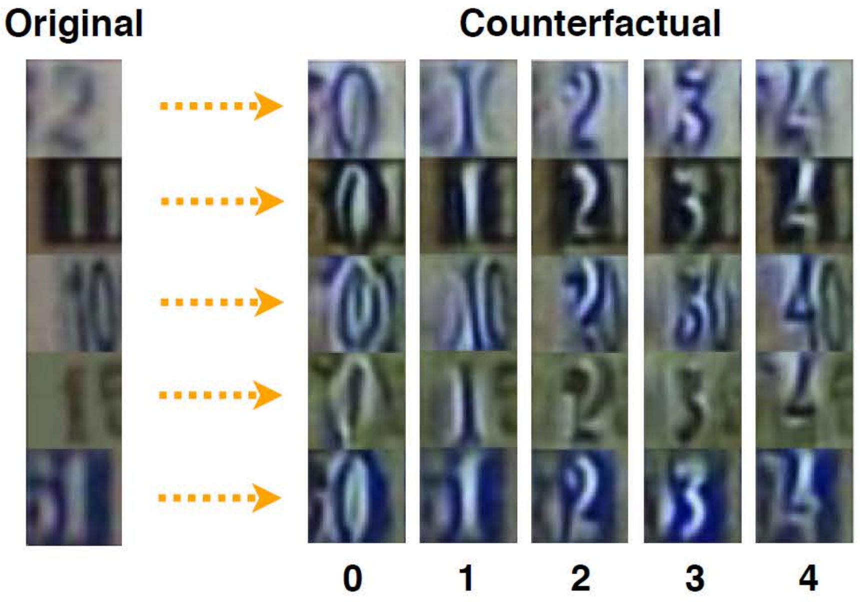
- Train Conditional GAN (DCGAN)
  - Conditioned on the K known classes
  - Input encoded as one-hot vector
- Train a standard K-class classifier
- Generate candidate “aliens”
  - Feed mixture vectors as input (0,0,0.5,0.5,0,…,0)
  - If the classifier is confidence, then discard the candidate
- Now train OpenMax network with $K + 1$ classes
  - REJECT if either the classifier predicts the “alien” class $(K + 1)$ or class 0

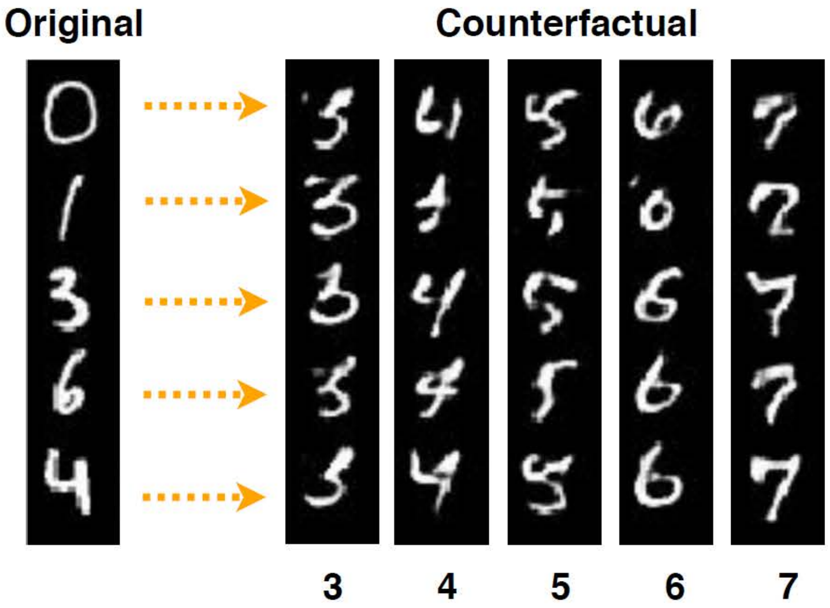
Counterfactual Examples

(Neal, Olson, Fern, Wong, Li, arXiv 2018)

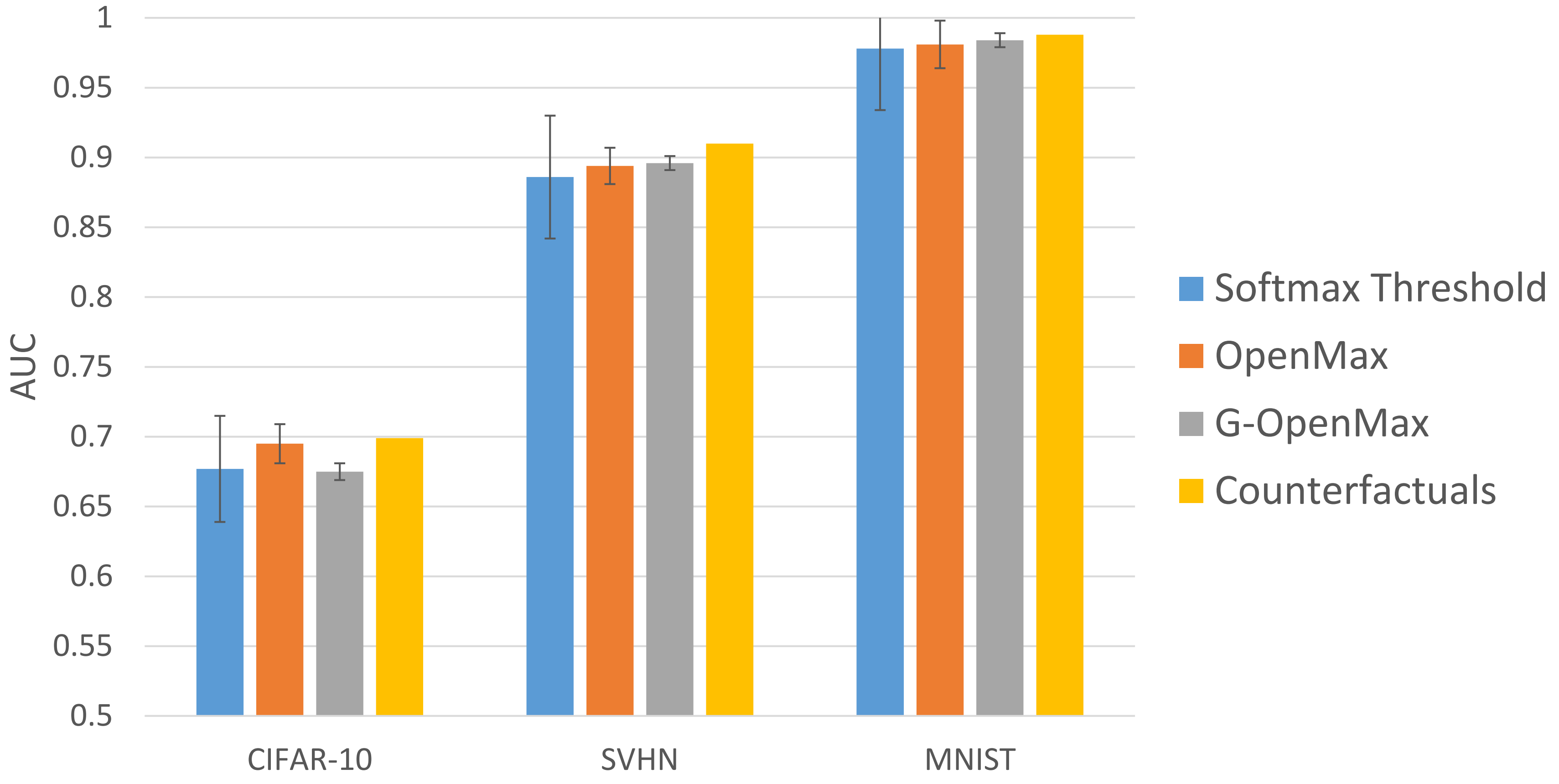
- Generate an example x that resembles target class K as much as possible but lies on the "other side" of the decision boundary separating the known and unknown classes
- Let E be an encoder, G the generator of a DCGAN, and C_K be a K-class classifier with softmax output
- Let $C_K(x)_k$ be the logit of class k
- $z^* = \min_z \|z - E(x)\|^2 + \log\left(1 + \sum_{k=1}^{K} C_K\big(G(z)\big)_k\right)$

Example Counterfactual Images



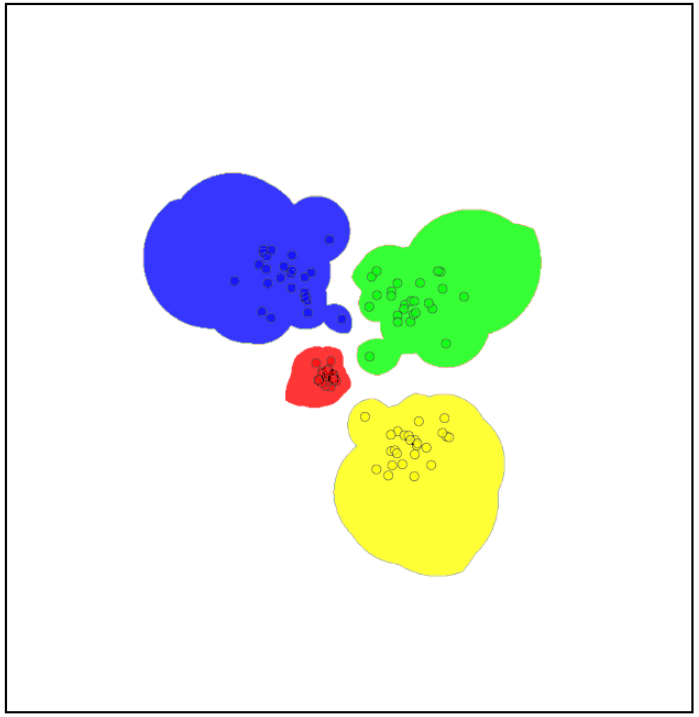


Results



1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
AUC
CIFAR-10
SVHN
MNIST
Softmax Threshold
OpenMax
G-OpenMax
Counterfactuals

Other Methods

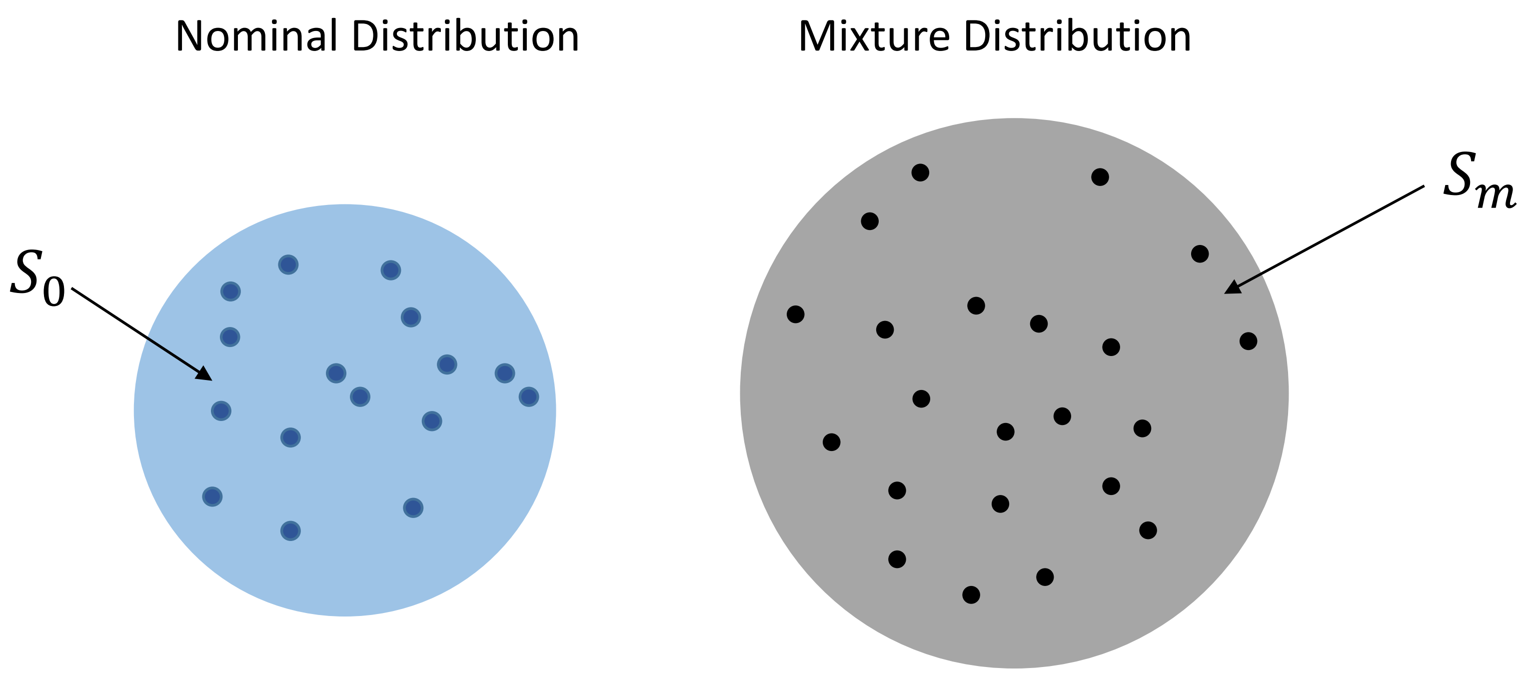


- OSNN: Nearest Neighbor Distance Ratio (Mendes-Junior et al. MLJ 2017)
  - Let $n_1 = (x_1, y_1)$ be the nearest neighbor to x_q
  - Let $n_2 = (x_2, y_2)$ be the nearest neighbor to x_q whose class $y_2 \neq y_1$
  - $ratio = \frac{d(x_q, x_1)}{d(x_q, x_2)}$
  - If $ratio > \tau$ then REJECT
  - Else classify as $\hat{y} = y_1$
- ODIN (Liang, Li, Srikant, ICLR 2018)
  - Tune softmax temperature T
  - Let $\hat{S} := \hat{p}(\hat{y} | x_q; T)$ be the softmax score of the input query
  - Let $\tilde{x}_q = x - \epsilon[sgn(-\nabla_x \log \hat{S})]$
  - Let $\tilde{S} := \hat{p}(\hat{y} | \tilde{x}_q; T)$ be the softmax score of the perturbed instance
  - If $\tilde{S} < \tau$ REJECT else classify as $\hat{y}$

Another Resource: Unlabeled Data

- Da, Yu, Zhou (2014) "Learning with Augmented Class by Exploiting Unlabeled Data"
  - Formulate a kind of semi-supervised learning problem to find a decision boundary separating each known from the unknown classes
- Liu et al (2018). Use unlabeled data to set the rejection threshold

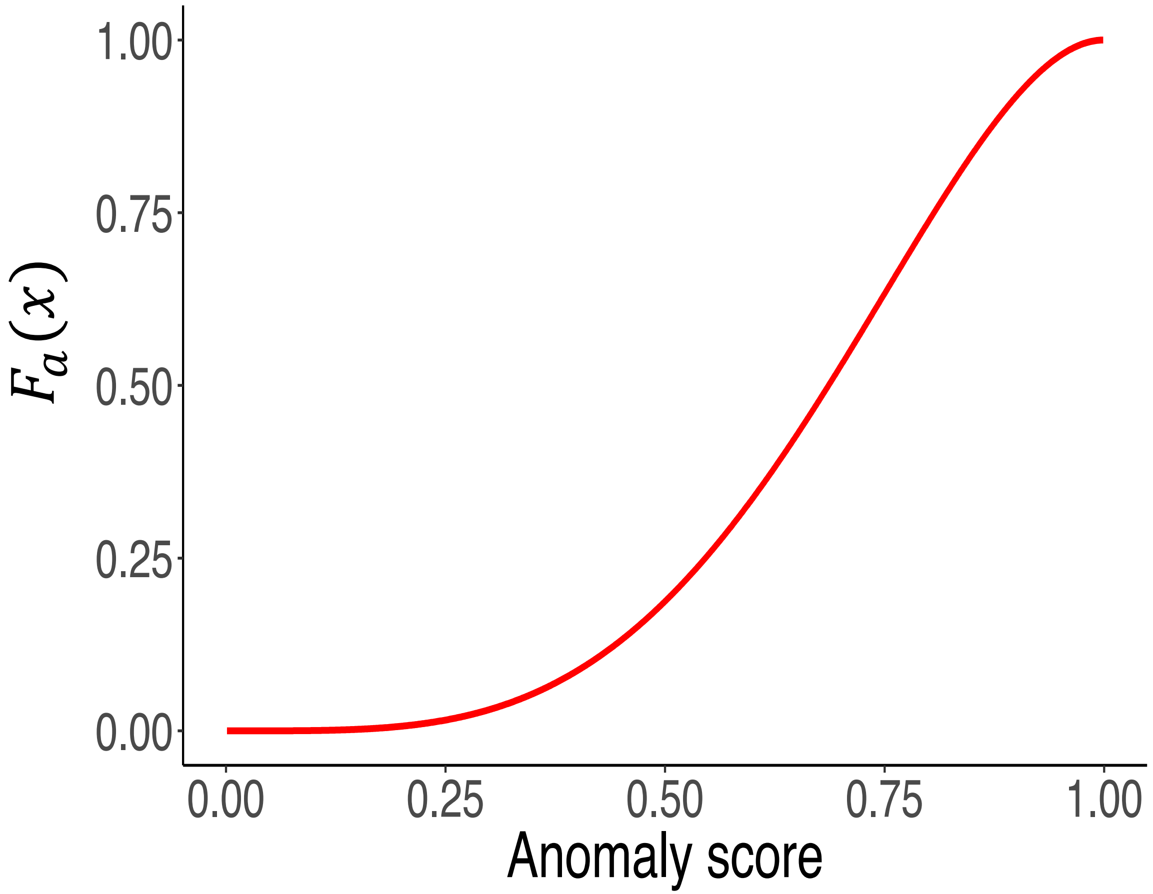
Obtaining Theoretical Guarantees



Proportion of Aliens = α

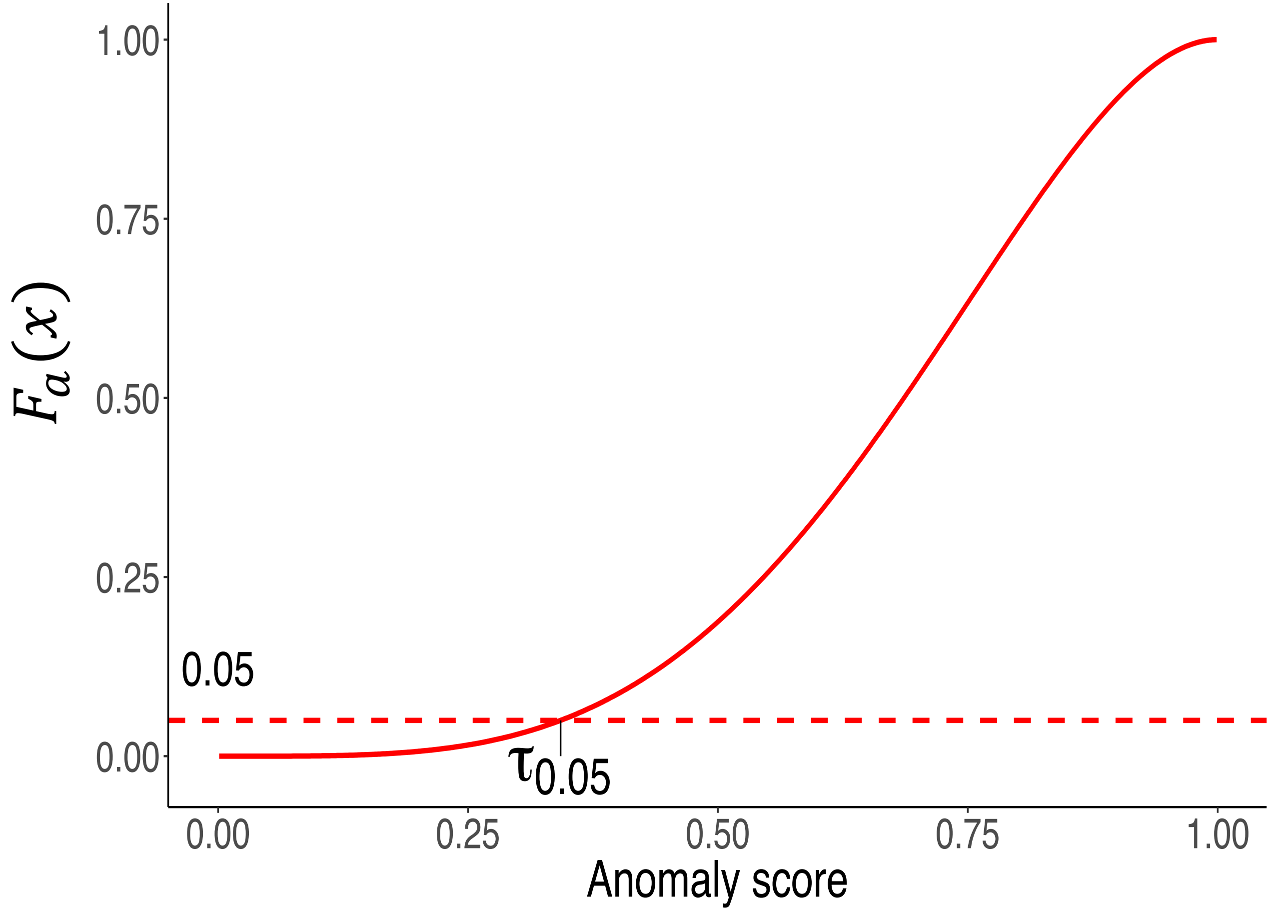
$$P_m = (1-\alpha)P_0 + \alpha P_a$$

Cumulative CDF of Alien Anomaly Scores: F_a



Want to have recall $= 1 - q$

Choosing τ for target quantile q



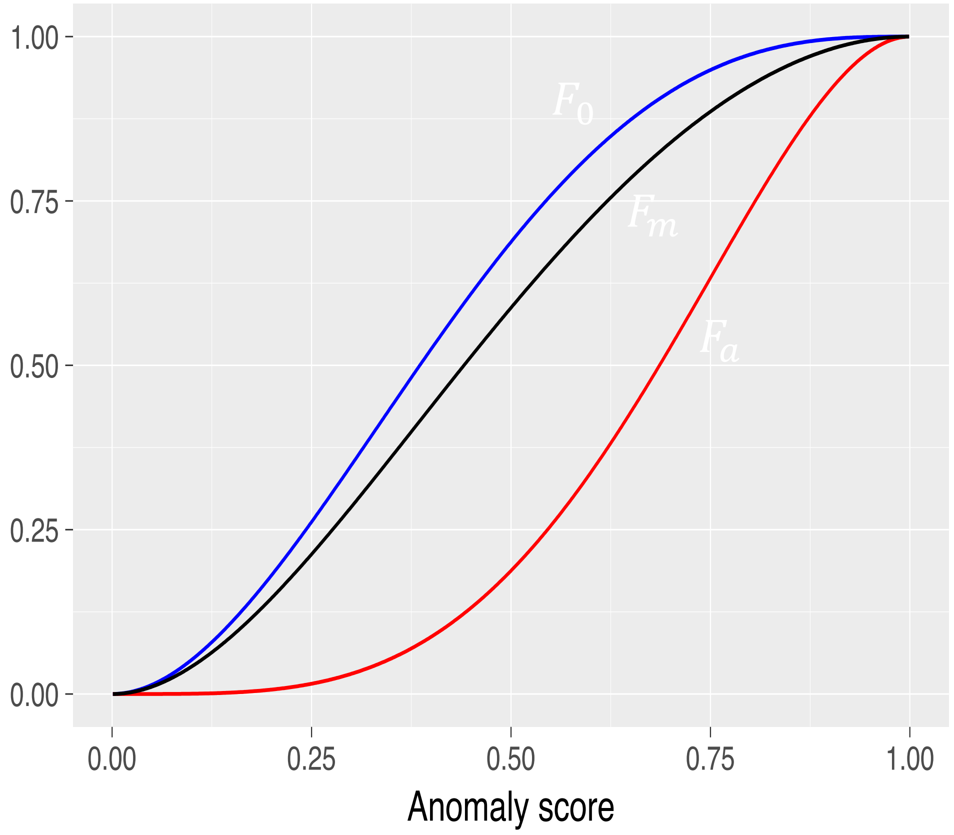
$$q = 0.05$$

$$P_m = (1-\alpha)P_0 + \alpha P_a$$

implies that

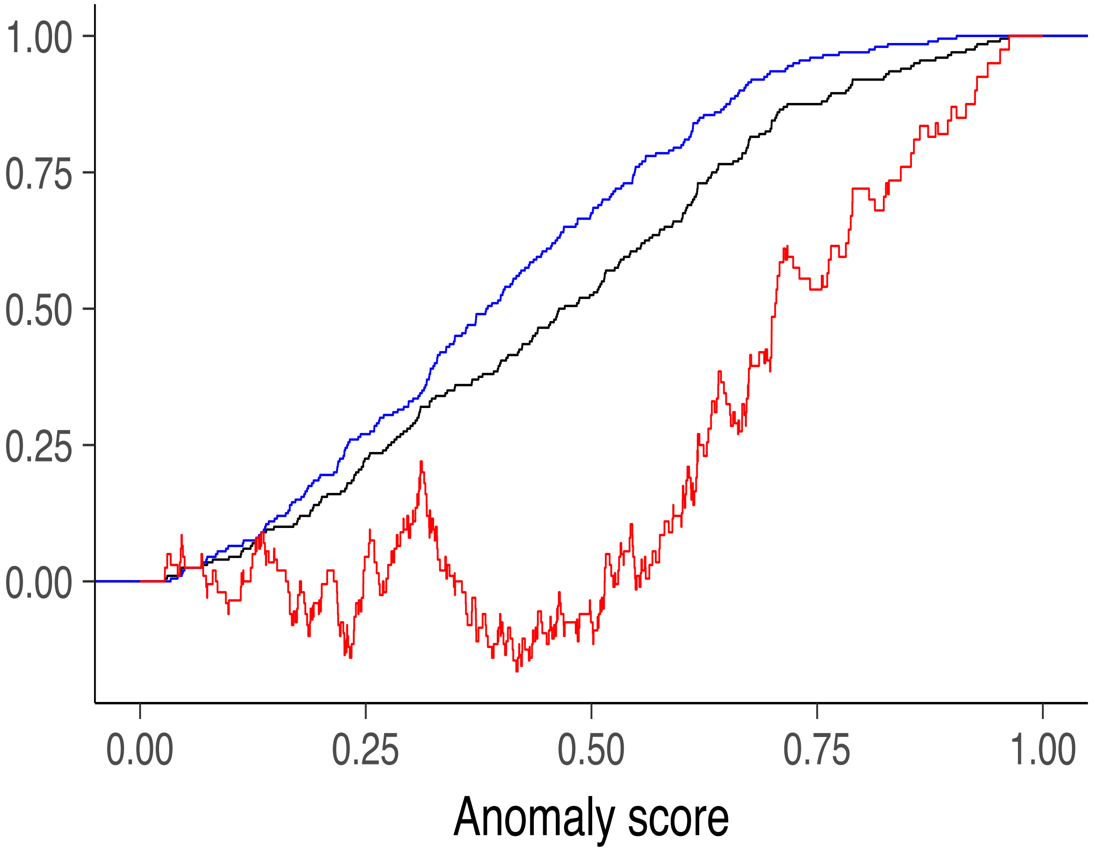
$$F_m(x) = (1-\alpha)F_0(x) + \alpha F_a(x)$$

CDFs of Nominal, Mixture, and Alien Anomaly Scores



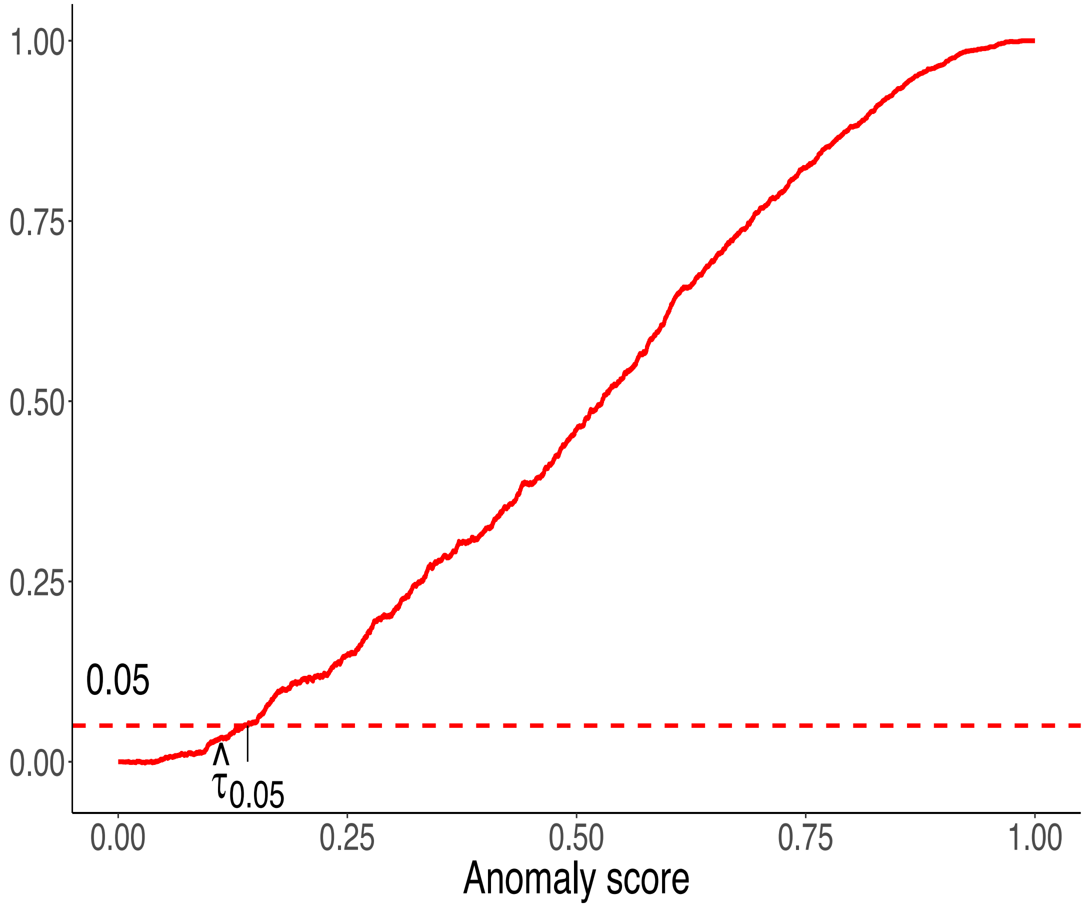
$$F_a(x) = \frac{F_m(x) - (1-\alpha)F_0(x)}{\alpha}$$

What We Have Are Empirical CDFs



$$\hat{F}_a(x) = \frac{\hat{F}_m(x) - (1-\alpha)\hat{F}_0(x)}{\alpha}$$

We Use the Empirical Estimate $\hat{\tau}_{0.05}$



EstimateTau($\boldsymbol{S_0}, \boldsymbol{S_m}, \boldsymbol{q}, \boldsymbol{\alpha}$)

1: Anomaly scores of S_0: $x_1, x_2, \cdots, x_k$

2: Anomaly scores of S_m: $y_1, y_2, \cdots, y_m$

3: Compute empirical CDFs $\hat{F}_0$ and $\hat{F}_m$.

4: Calculate $\hat{F}_a$ using

$$\hat{F}_a(x) = \frac{\hat{F}_m(x) - (1-\alpha)\hat{F}_0(x)}{\alpha}.$$

5: Output detection threshold

$$\hat{\tau}_q = \max_{u \in S} \hat{F}_a(u) \leq q,$$

where $S = \{x_1, x_2, \cdots, x_k, y_1, y_2, \cdots, y_m\}$.

Theoretical Guarantee

[Liu, Garrepalli, Fern, Dietterich, ICML 2018]

- Theorem: If

$$n > \frac{1}{2} \ln \frac{2}{1 - \sqrt{1 - \delta}} \left(\frac{1}{\epsilon}\right)^2 \left(\frac{2 - \alpha}{\alpha}\right)^2$$

then with probability $1 - \delta$ the alien detection rate will be at least $1 - (q + \epsilon)$

What if we don't know the exact value of α?

Def: We say that an anomaly detector is ***sufficient***, if the score CDFs satisfy

$$F_0(x) \geq F_a(x), \text{ for all } x \in \mathbb{R}.$$

Corollary 1: Replace α with α'

Assume F_0 and F_a sufficient, and continuous with convex support. $|S_0| = |S_m| = n$. For any $\epsilon \in (0, 1-q)$ and δ $\in (0,1)$, if

$$n \geq \frac{1}{2} \ln \frac{2}{1-\sqrt{1-\delta}} \left(\frac{1}{\epsilon}\right)^2 \left(\frac{2-\alpha'}{\alpha'}\right)^2,$$

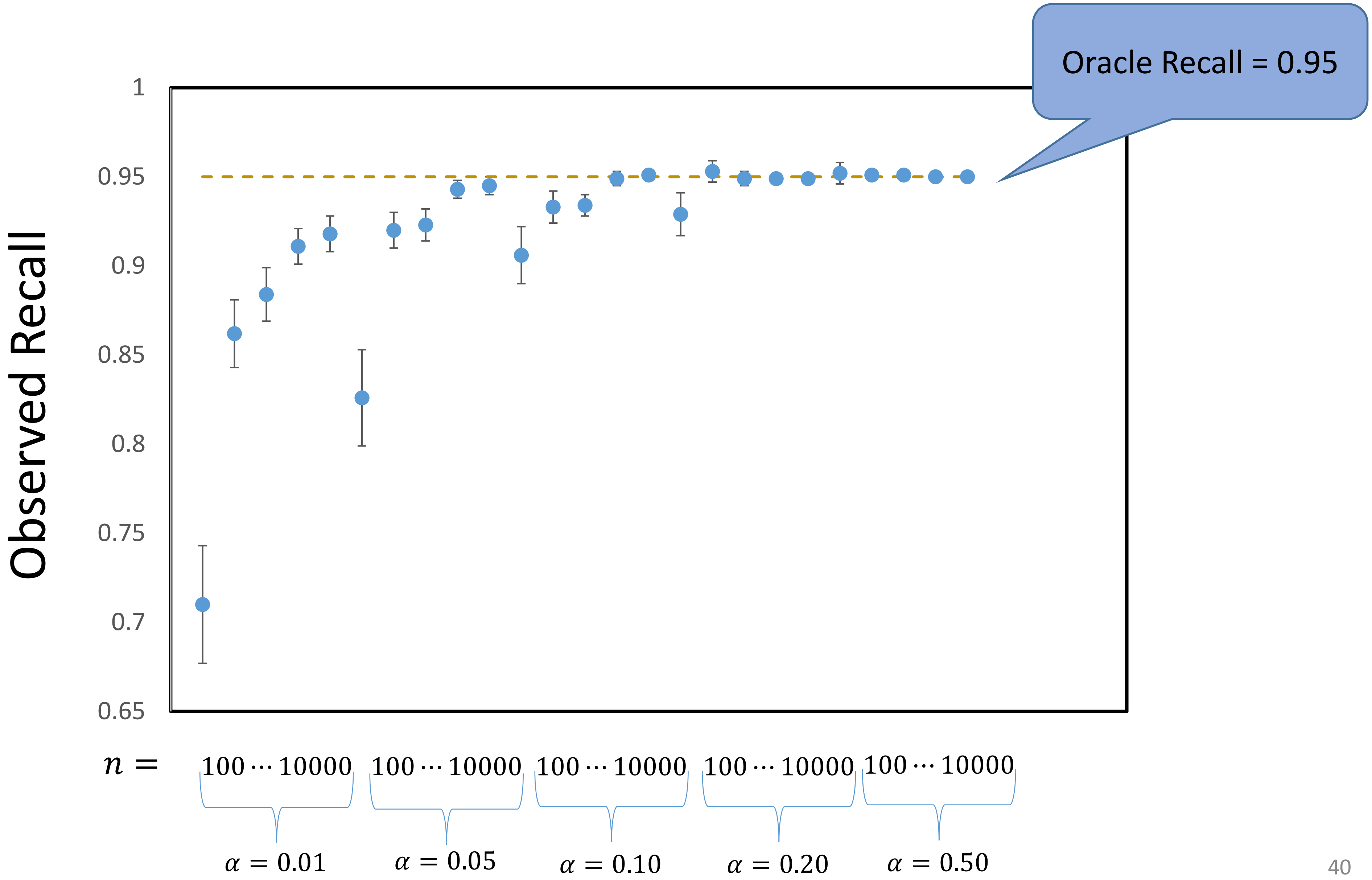
Algorithm 1 will return a threshold $\hat{\tau}_q$ that achieves an alien detection rate of at least $1-(q+\epsilon)$ with probability $1-\delta$

Note: $\hat{\tau}$ will be more conservative

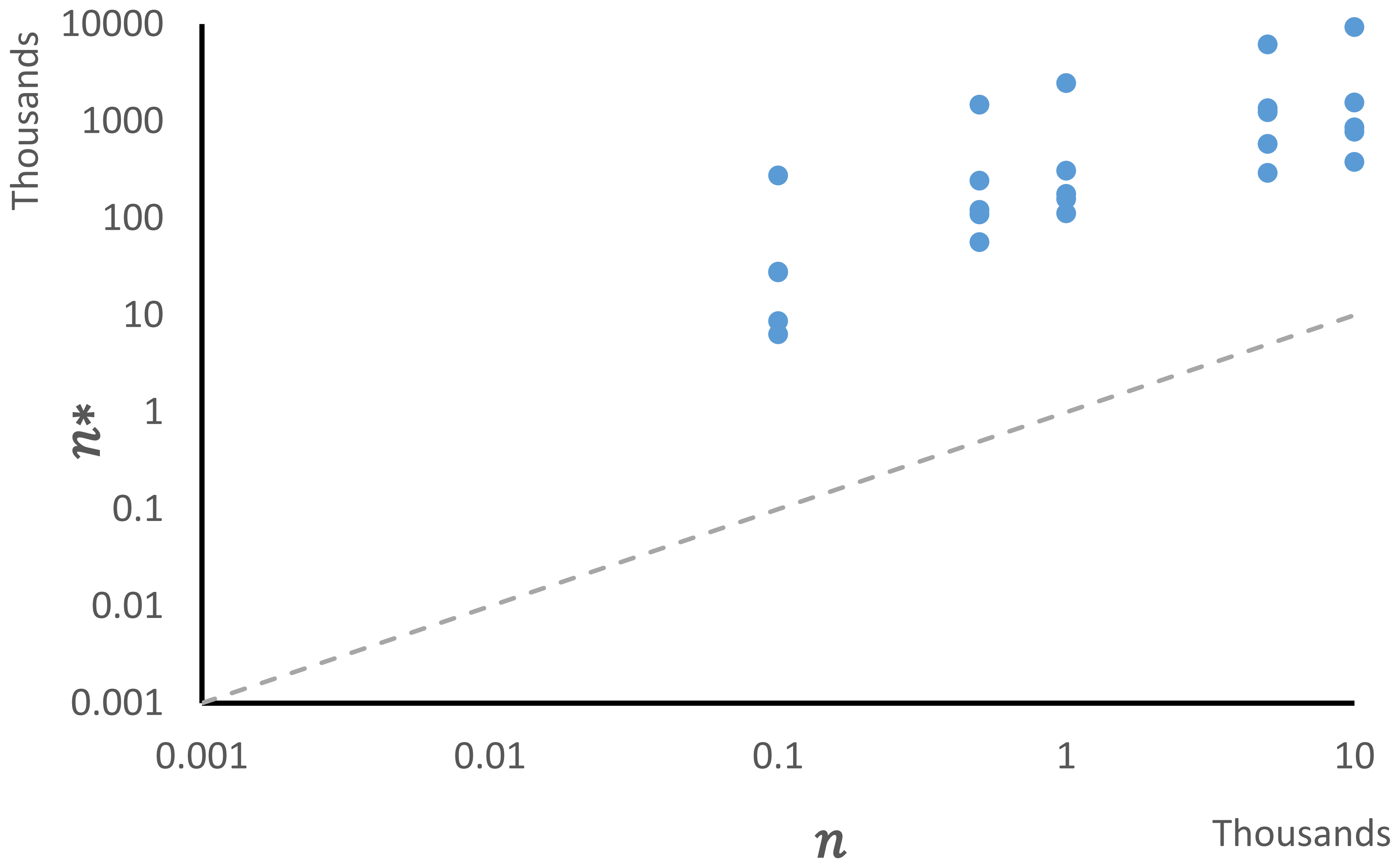
Four Experimental Questions

1. How accurate is our estimate of τ_q?
2. How loose is the bound on n?
3. How good are Recall and FAR in practice?
4. What is the impact of using $\alpha' > \alpha$?

Q1: How accurate is our estimate of τ_q?

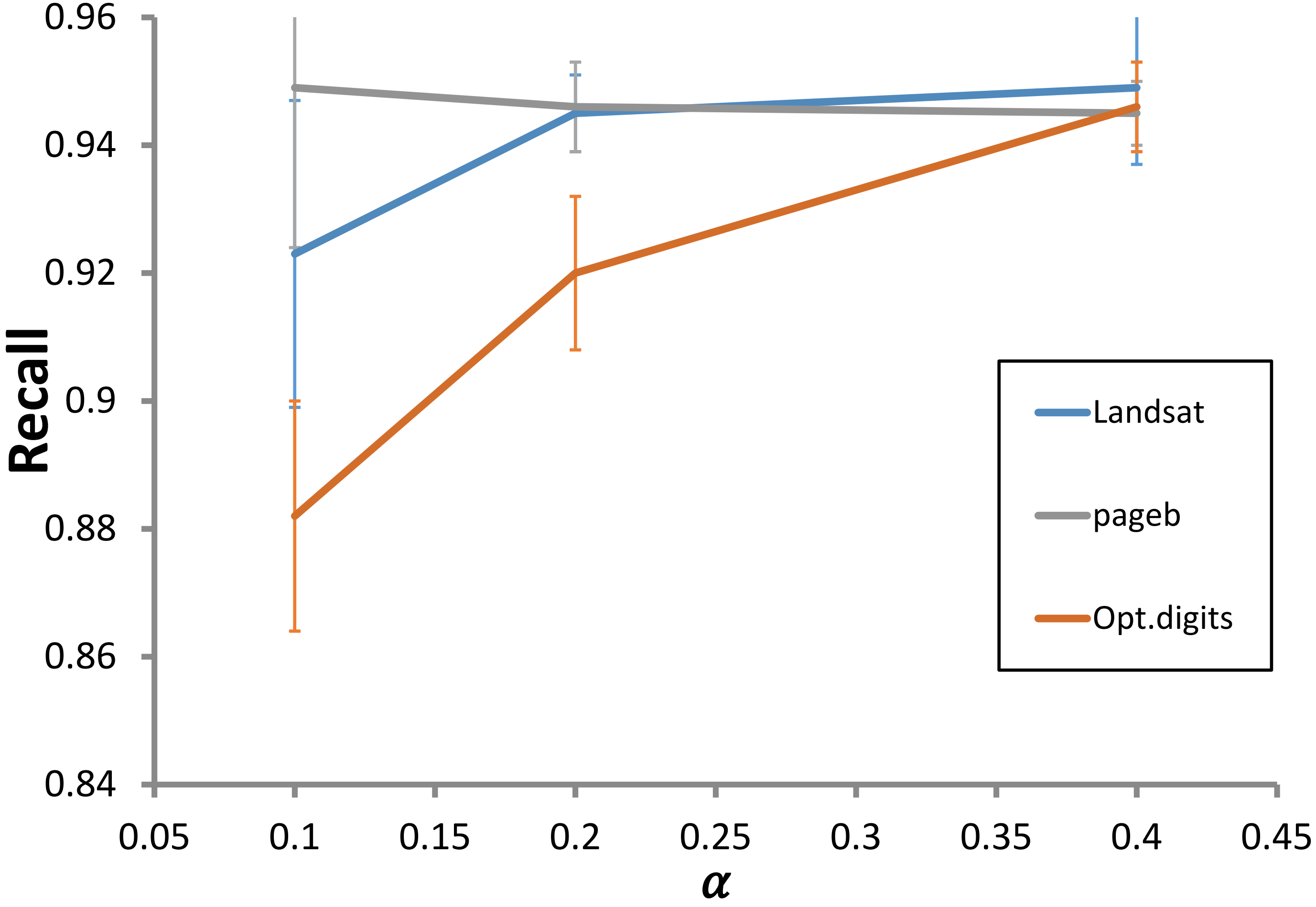


Q2: How loose is the bound on n?

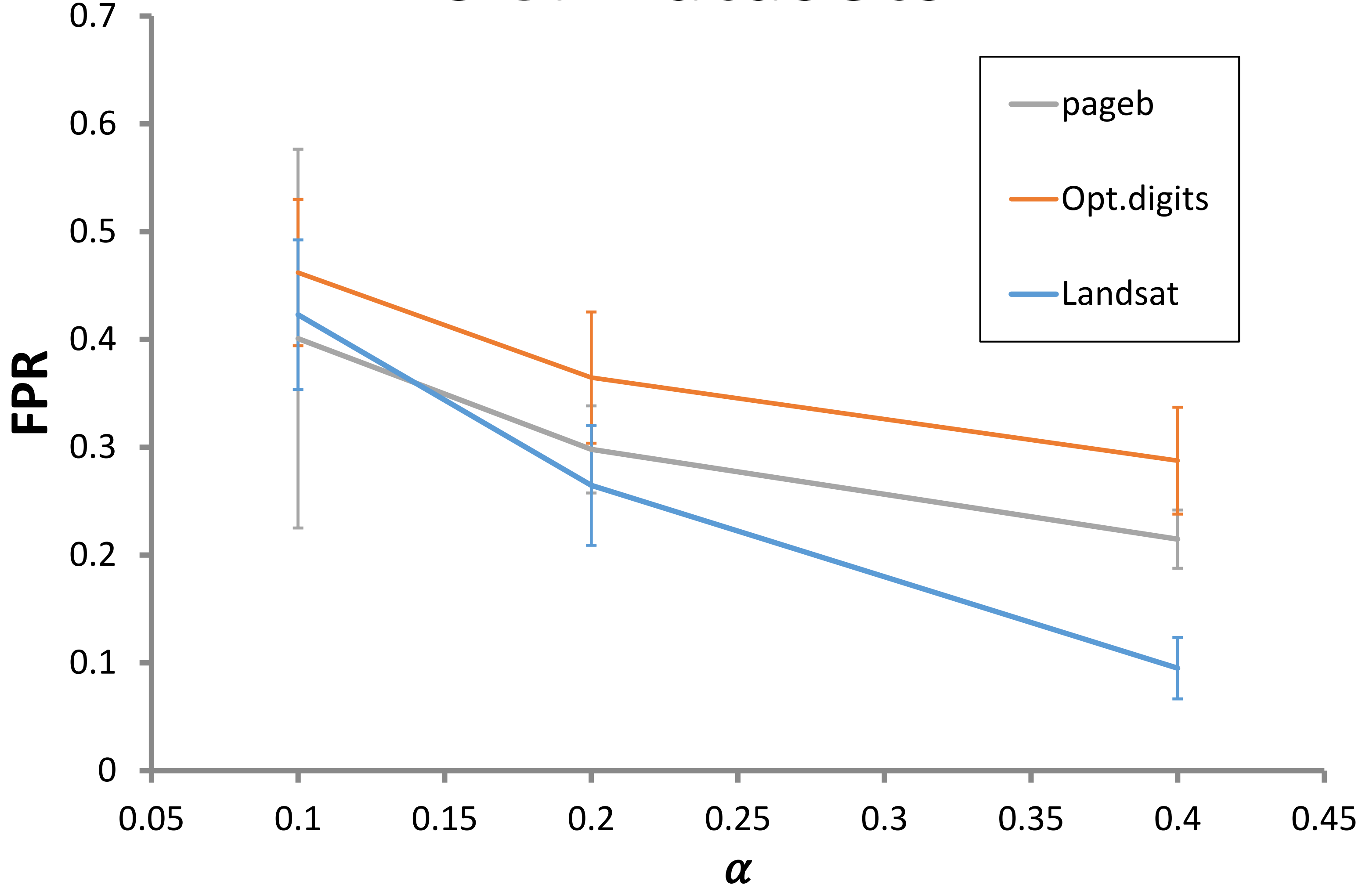


Q3: How good are Recall and FPR in practice? UCI Datasets

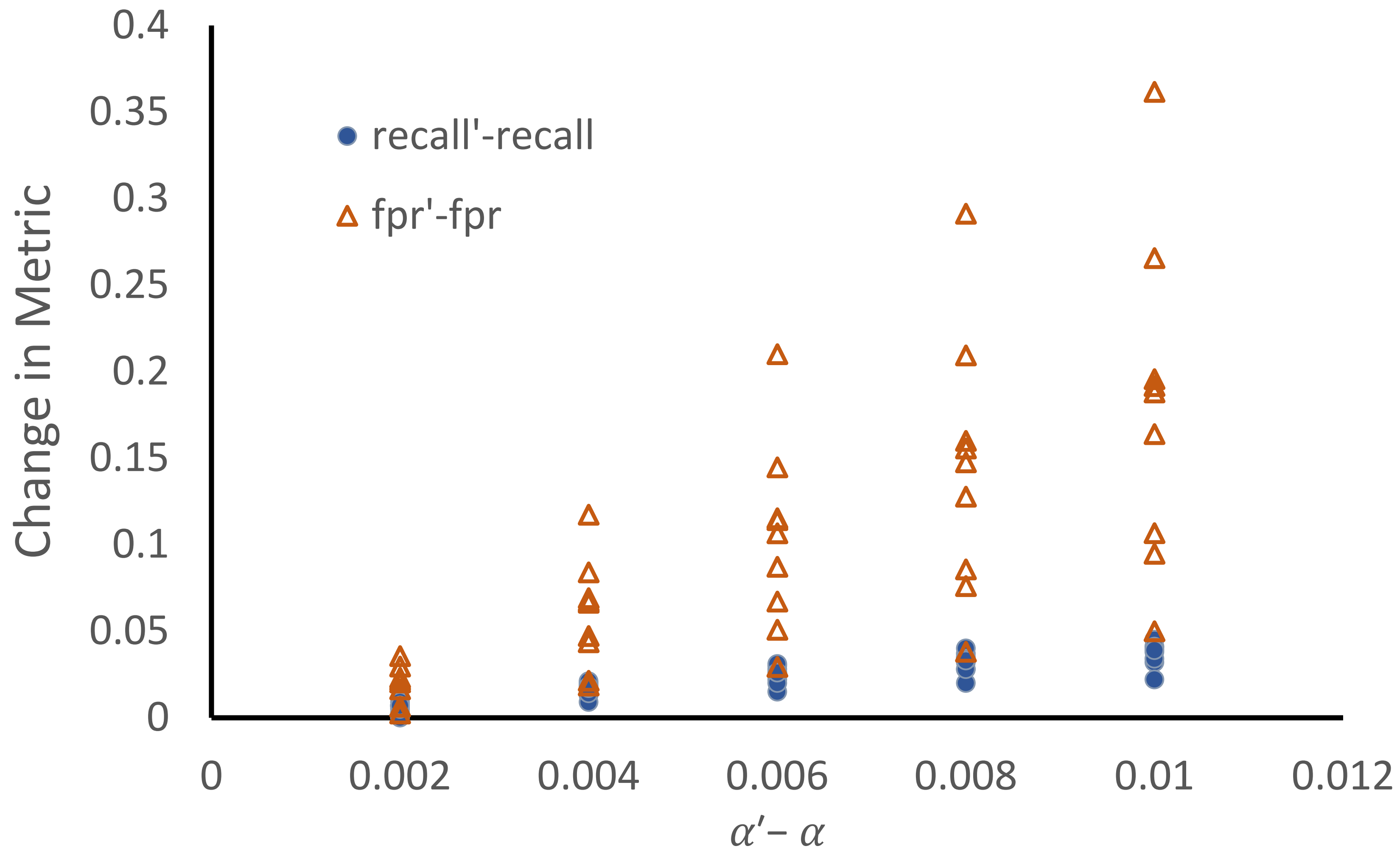
$q = 0.95$



Q3: How good are Recall and FPR in practice? UCI Datasets



Q4: What is the impact of using $\alpha' > \alpha$?



Assessment

- This area is mostly an empirical mess and lacks theory
- Our PAC result requires access to unlabeled data containing aliens
  - AND a tight upper bound on α

Next Lecture: Anomaly Detection

- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-Based Anomaly Detection. ACM Transactions on Knowledge Discovery from Data, 6(1), 1–39. http://doi.org/10.1145/2133360.2133363
- Emmott, A., Das, S., Dietterich, T., Fern, A., & Wong, W.-K. (2015). Systematic construction of anomaly detection benchmarks from real data. https://arxiv.org/abs/1503.01158
- Siddiqui, A., Fern, A., Dietterich, T. G., & Das, S. (2016). Finite Sample Complexity of Rare Pattern Anomaly Detection. In *Proceedings of UAI-2016* (p. 10). http://auai.org/uai2016/proceedings/papers/226.pdf

Citations

- Bendale, A., & Boult, T. (2016). Towards Open Set Deep Networks. In CVPR 2016 (pp. 1563–1572). http://doi.org/10.1109/CVPR.2016.173
- Bodesheim, P., Freytag, A., Rodner, E., Kemmler, M., Denzler, J. (2013) Kernel Null Space Methods for Novelty Detection. *CVPR 2013.*
- Da, Q., Yu, Y., Zhou, Z-H. (2014) Learning with Augmented Class by Exploiting Unlabeled Data, *AAAI 2014*.
- Ge, Z., Demyanov, S., Chen, A., Garnavi, B. (2017). Generative OpenMax for Multi-Class Open Set Classification. arXiv 1707.07418.
- Hassen, M., Chan, P., (2018). Learning a Neural-network-based Representation for Open Set Recognition. arXiv 1802.04365.
- Liang, S., Li, Y., Srikant, R. (2018). Enhancing the Reliability of Out-of-distribution Image Detection in Neural Networks. *ICLR 2018.*
- Liu, S., Garrepalli, R., Dietterich, T. G., Fern, A., & Hendrycks, D. (2018). Open Category Detection with PAC Guarantees. *Proceedings of the 35th International Conference on Machine Learning, PMLR*, *80*, 3169–3178. http://proceedings.mlr.press/v80/liu18e.html
- Mendes-Júnior, P., de Souza, R.,Werneck, R., Stein, B. V., Pazinato, D. V., de Almeida, W. R. (2017). Nearest neighbors distance ratio open-set classifier. *Machine Learning* 106: 359–386.
- Neal, L., Olson, M., Fern, A., Wong, W-K., Li, F. (2018). Open Set Learning with Counterfactual Images. Proceedings of the European Conference on Computer Vision (ECCV 2018).
- Salakhutdinov, R., Torralba, A., Tenenbaum, J. (2011) Learning to Share Visual Appearance for Multiclass Object Detection. *CVPR 2011.*
- Shafaei, A., Schmidt, M., & Little, J. (2018). Does Your Model Know the Digit 6 Is Not a Cat? A Less Biased Evaluation of “Outlier” Detectors. arXiv 1809.04729