Efficient top rank optimization with gradient boosting for supervised anomaly detection

By

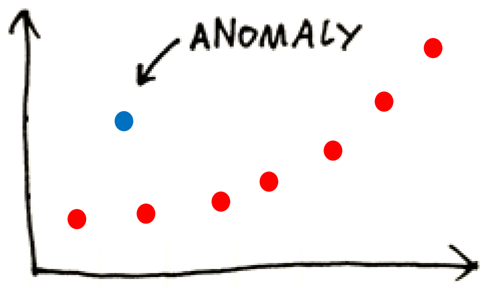
Jordan Fréry, Amaury Habrard, Marc Sebban, Olivier Caelen and Liyun He-Guelton

Anomaly detection

Find the data that do not conform to the normal behavior.

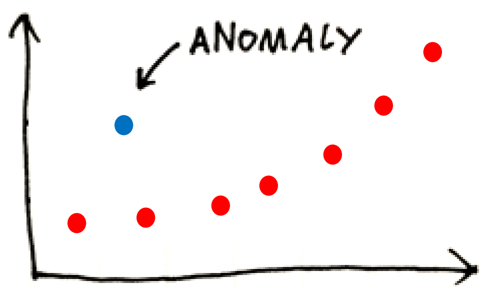
Anomaly detection

Find the data that do not conform to the normal behavior.



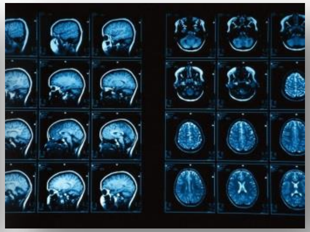
Anomaly detection

Find the data that do not conform to the normal behavior.



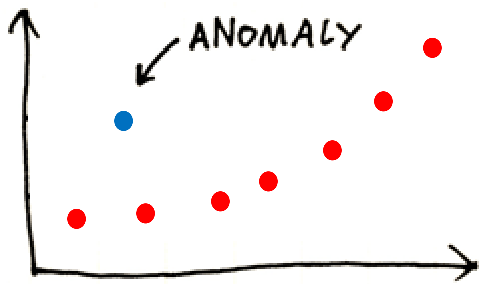
Applications:

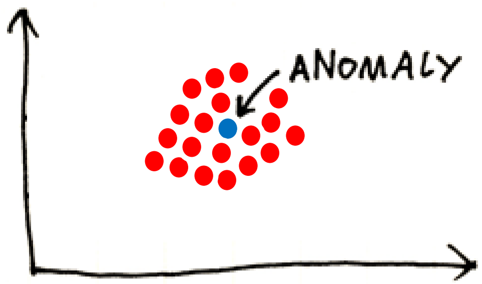
- Health care
- Intrusion detection in cyber security



Anomaly detection

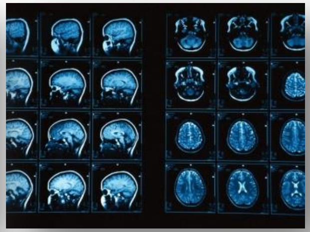
Find the data that do not conform to the normal behavior.





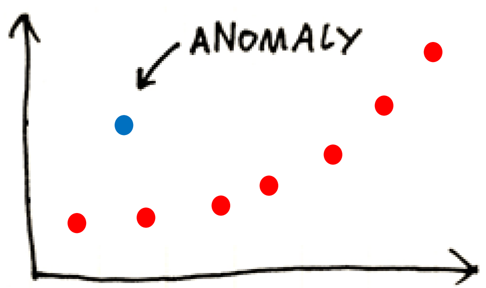
Applications:

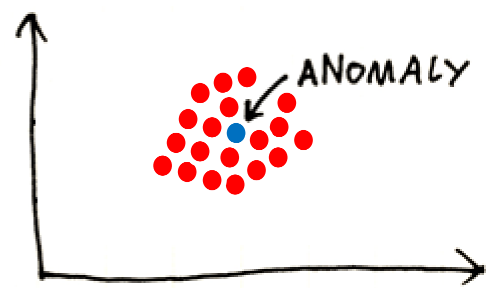
- Health care
- Intrusion detection in cyber security



Anomaly detection

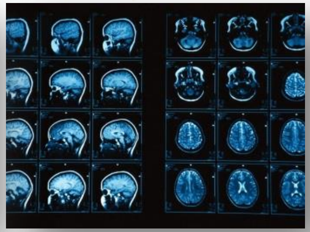
Find the data that do not conform to the normal behavior.





Applications:

- Health care
- Intrusion detection in cyber security
- Fraud detection
  - Insurance
  - Credit card transactions



Supervised learning

- Goal is to find a function $f : X \rightarrow Y$

Supervised learning

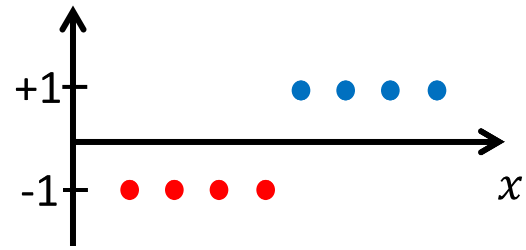
- Goal is to find a function $f : X \to Y$
- We assume that there is an unknown joint distribution P over $X \times Y$

Supervised learning

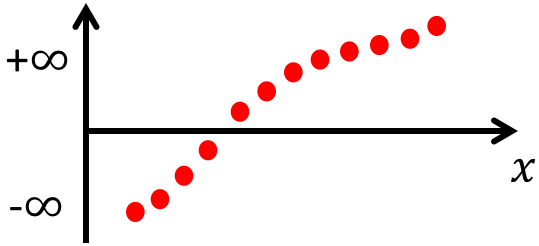
- Goal is to find a function $f : X \ \to Y$
- We assume that there is an unknown joint distribution P over $X \times Y$
- We have a training set of M examples $\{x_i, y_i\}_{n=1}^{M} \in (X \times Y)^M$ i.i.d. according to P

Supervised learning

- Goal is to find a function $f : X \to Y$
- We assume that there is an unknown joint distribution P over $X \times Y$
- We have a training set of M examples $\{x_i, y_i\}_{n=1}^{M} \in (X \times Y)^M$ i.i.d. according to P
- Classification: discrete y



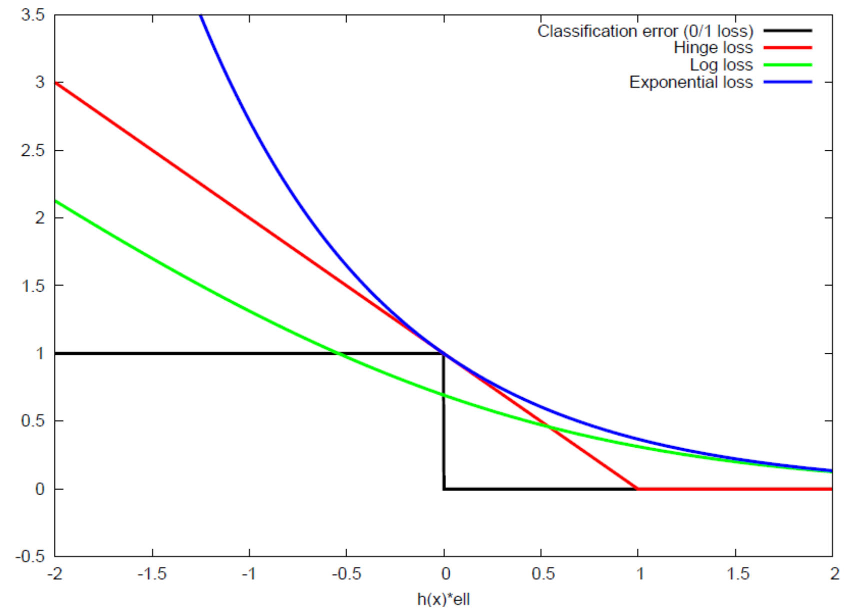
- Regression: continuous y



Loss function and risk minimization

Loss function and risk minimization

- Loss function $L(f(x), y)$: agreement between prediction $f(x)$ and desired output y.

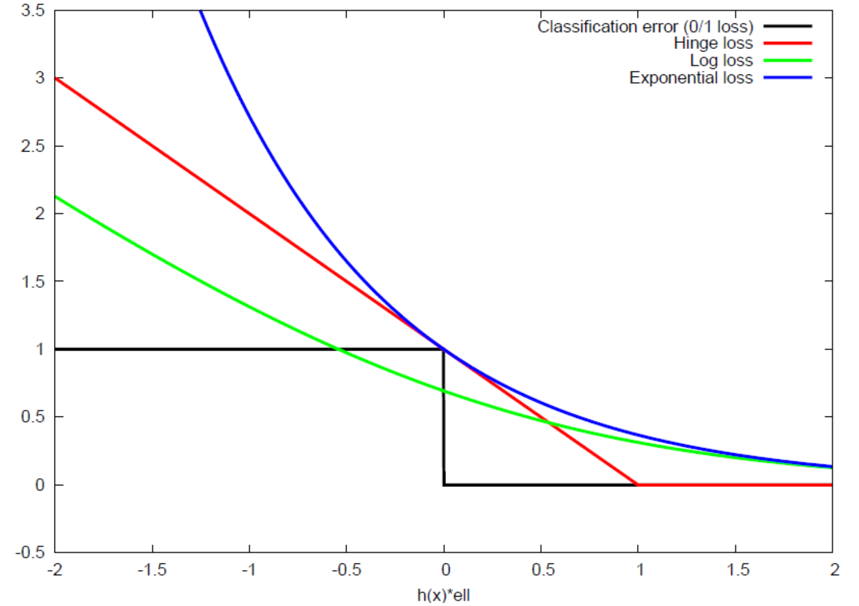


Example of different loss functions

Loss function and risk minimization

- Loss function $L(f(x), y)$: agreement between prediction $f(x)$ and desired output y.

- Common loss functions are accuracy based



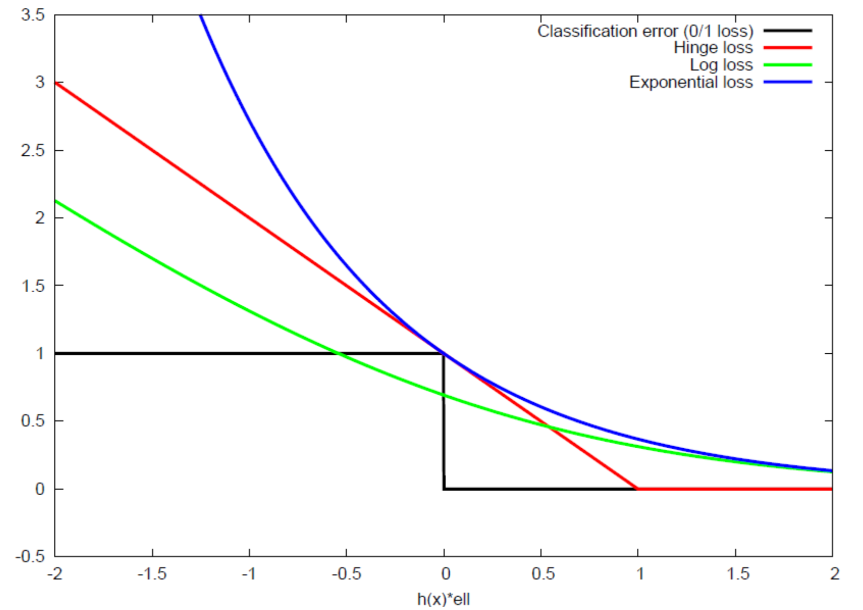
Example of different loss functions

Loss function and risk minimization

- Loss function $L(f(x), y)$: agreement between prediction $f(x)$ and desired output y.

- Common loss functions are accuracy based

- Anomaly detection can be presented as a binary problem with $N \gg P$



Example of different loss functions

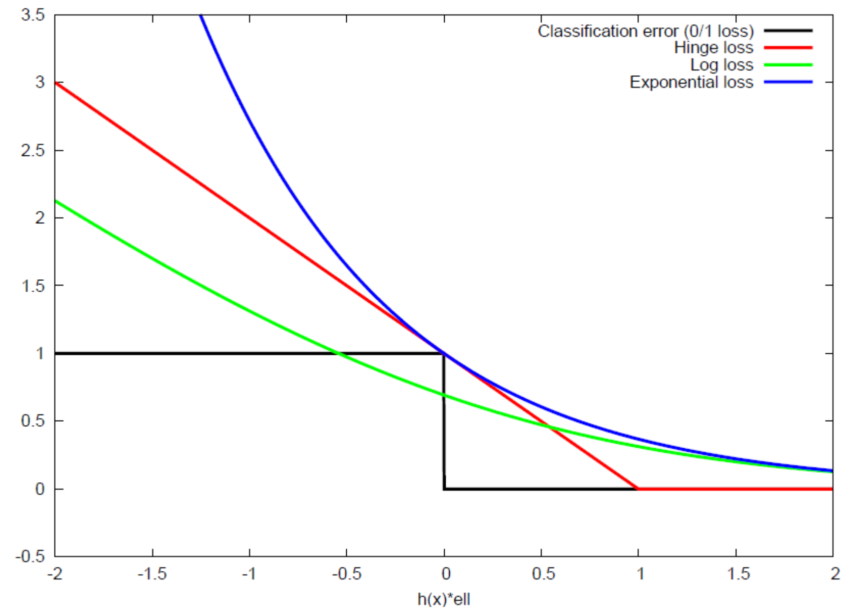
Loss function and risk minimization

- Loss function $L(f(x), y)$: agreement between prediction $f(x)$ and desired output y.

- Common loss functions are accuracy based

- Anomaly detection can be presented as a binary problem with $N \gg P$

- In anomaly detection, we would rather look at
  - $Precision$
  - $Recall$
  - $F_\beta\ score$

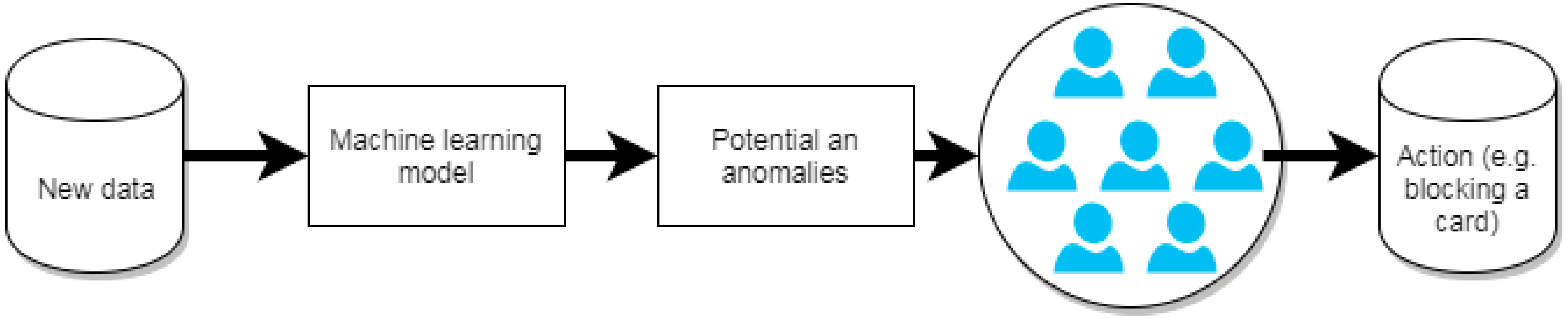


Example of different loss functions

Pitfall of the classification approach

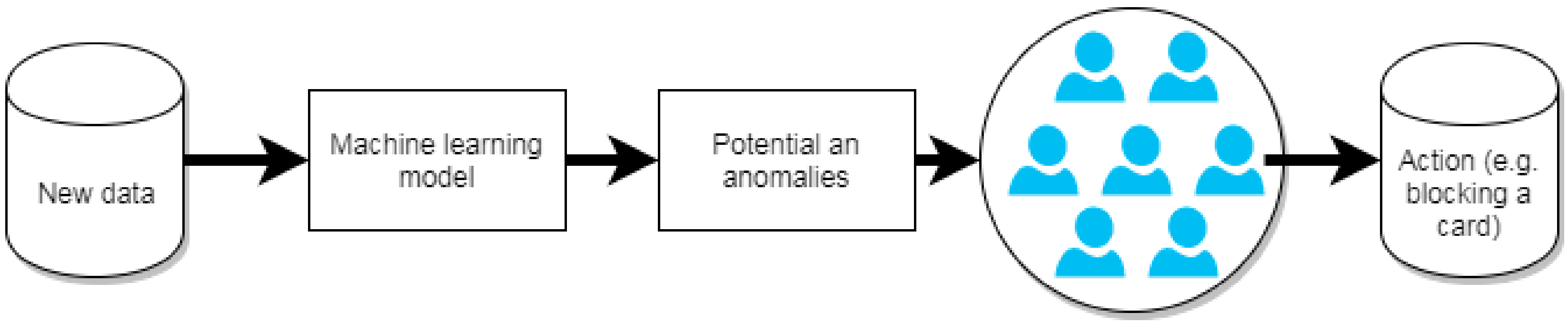
Pitfall of the classification approach

- Experts often need to assess the potential anomalies



Pitfall of the classification approach

- Experts often need to assess the potential anomalies



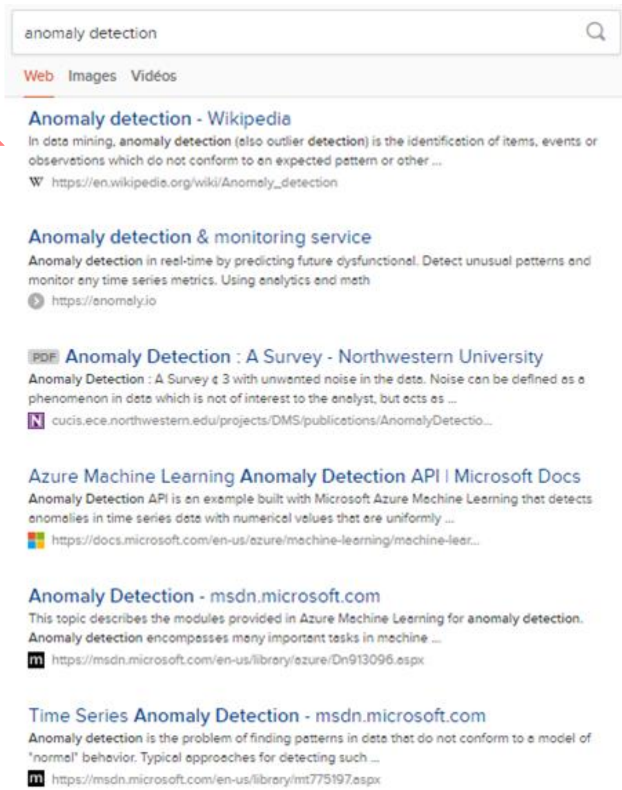
Use another approach than classification: learning to rank

Learning to rank

Learning to rank

Predicted most relevant
links for a query:

Relevance rank predicted

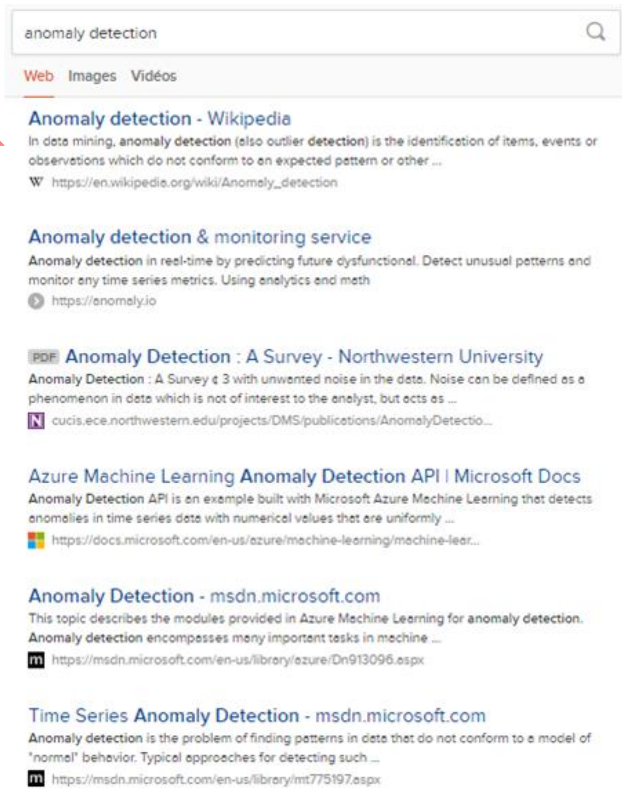


+ millions more links

Learning to rank

Predicted most relevant links for a query:

Relevance rank predicted



+ millions more links

Predicted most probable fraudulent transaction over a day:

Fraudulent rank predicted

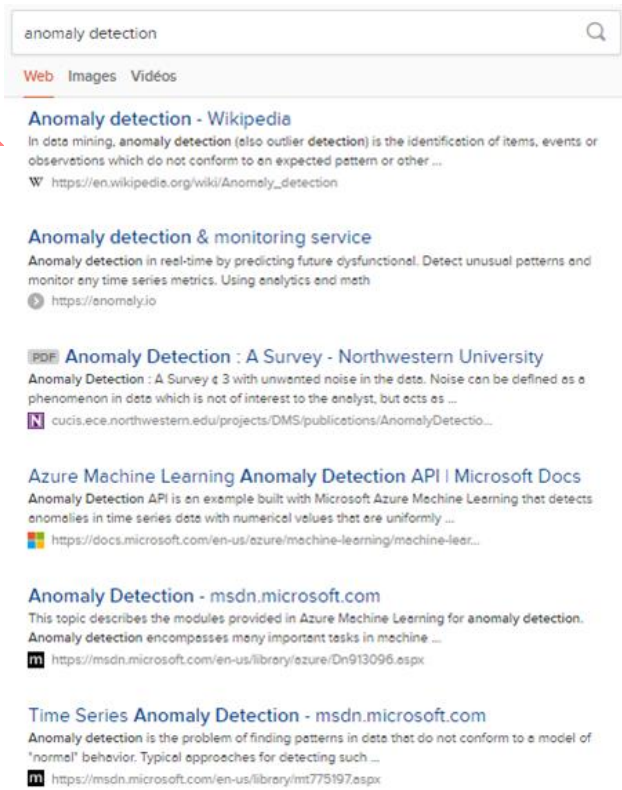
Card ID	Date	Amount	Shop.
120983	21/09/2017	23€	20193
328903	21/09/2017	3€	29103
328032	21/09/2017	14.2€	9023
390293	21/09/2017	11€	124
182393	21/09/2017	110€	202
432445	21/09/2017	43€	20193
645367	21/09/2017	4€	1089
887644	21/09/2017	34.3€	230
257546	21/09/2017	3.2€	399
655655	21/09/2017	40€	394
356578	21/09/2017	49.99€	12
884733	21/09/2017	73.99€	9000
...	...	...	...

+ millions more transactions

Learning to rank

Predicted most relevant links for a query:

Relevance rank predicted



+ millions more links

Predicted most probable fraudulent transaction over a day:

Fraudulent rank predicted

Card ID	Date	Amount	Shop.
120983	21/09/2017	23€	20193
328903	21/09/2017	3€	29103
328032	21/09/2017	14.2€	9023
390293	21/09/2017	11€	124
182393	21/09/2017	110€	202
432445	21/09/2017	43€	20193
645367	21/09/2017	4€	1089
887644	21/09/2017	34.3€	230
257546	21/09/2017	3.2€	399
655655	21/09/2017	40€	394
356578	21/09/2017	49.99€	12
884733	21/09/2017	73.99€	9000
...	...	...	...

+ millions more transactions

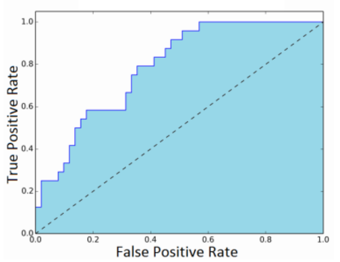
- How do we assess the quality of a ranked list?

Evaluation metrics for ranking

Evaluation metrics for ranking

- Area under the ROC curve

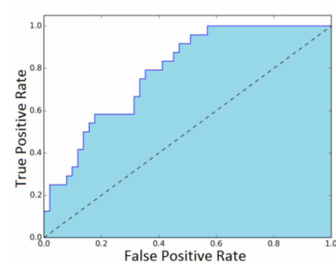
$$AUCROC = \frac{1}{PN}\sum_{i=1}^{P}\sum_{j=1}^{N} \mathrm{I}\left(f(x_i) > f\left(x_j\right)\right)$$

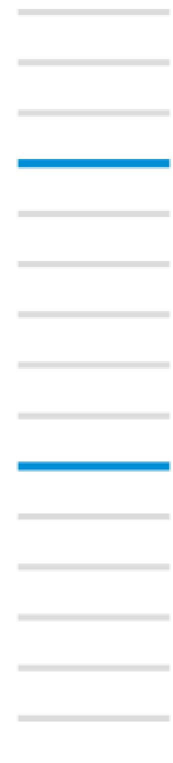


Evaluation metrics for ranking

- Area under the ROC curve

$$AUCROC = \frac{1}{PN}\sum_{i=1}^{P}\sum_{j=1}^{N} \mathrm{I}\left(f(x_i) > f(x_j)\right)$$



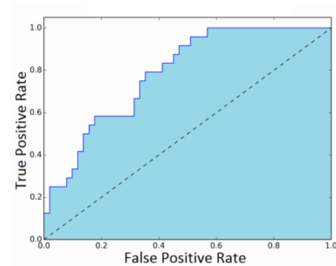


Burges 2010

Evaluation metrics for ranking

- Area under the ROC curve

$$AUCROC = \frac{1}{PN}\sum_{i=1}^{P}\sum_{j=1}^{N} \mathrm{I}\left(f(x_i) > f(x_j)\right)$$

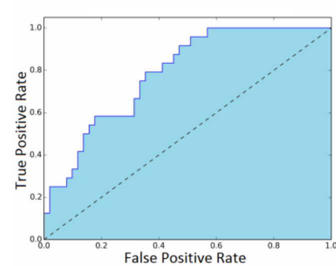


Burges 2010

Evaluation metrics for ranking

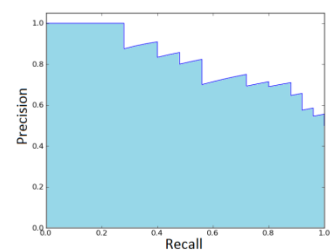
- Area under the ROC curve

$$AUCROC = \frac{1}{PN}\sum_{i=1}^{P}\sum_{j=1}^{N} \mathrm{I}\left(f(x_i) > f(x_j)\right)$$



- Average precision (AUCPR)

$$AP = \frac{1}{P}\sum_{i=1}^{P} precision@k$$

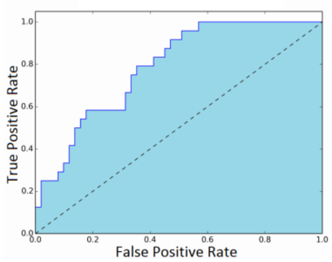


Burges 2010

Evaluation metrics for ranking

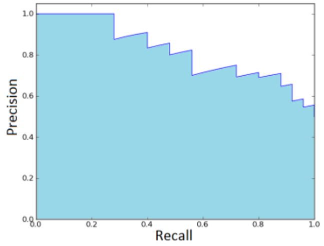
- Area under the ROC curve

$$AUCROC = \frac{1}{PN}\sum_{i=1}^{P}\sum_{j=1}^{N} \mathrm{I}\left(f(x_i) > f(x_j)\right)$$



- Average precision (AUCPR)

$$AP = \frac{1}{P}\sum_{i=1}^{P} precision@k$$

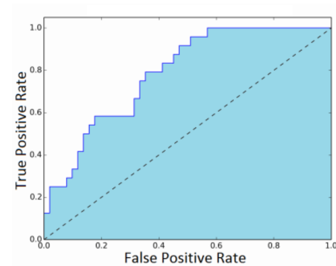


Burges 2010

Evaluation metrics for ranking

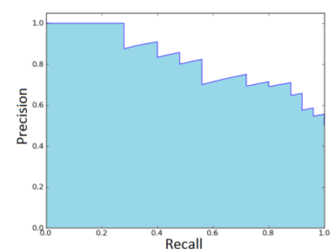
- Area under the ROC curve

$$AUCROC = \frac{1}{PN}\sum_{i=1}^{P}\sum_{j=1}^{N} \mathrm{I}\left(f(x_i) > f(x_j)\right)$$



- Average precision (AUCPR)

$$AP = \frac{1}{P}\sum_{i=1}^{P} precision@k$$



Burges 2010

AP is better suited to our problem. However, it is not differentiable.

Smooth approximation of AP

Smooth approximation of AP

- Average precision

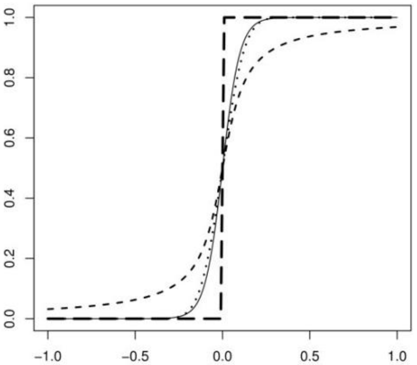
$$AP = \frac{1}{P}\sum_{i=1}^{P}\frac{\sum_{j=1}^{P}\mathrm{I}\left(f(x_i) < f\left(x_j\right)\right)}{\sum_{h=1}^{M}\mathrm{I}\left(f(x_h) < f(x_i)\right)}$$

Smooth approximation of AP

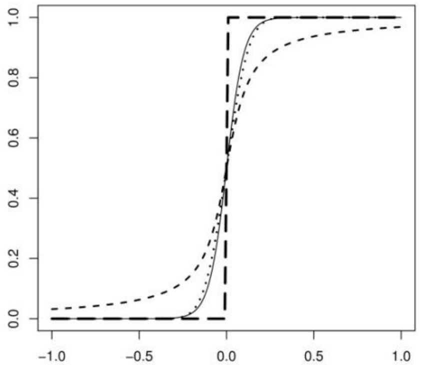
- Average precision

$$AP = \frac{1}{P}\sum_{i=1}^{P}\frac{\sum_{j=1}^{P}\mathrm{I}\left(f(x_i) < f(x_j)\right)}{\sum_{h=1}^{M}\mathrm{I}\left(f(x_h) < f(x_i)\right)}$$

- Sigmoid



Smooth approximation of AP



- Average precision

$$AP = \frac{1}{P}\sum_{i=1}^{P}\frac{\sum_{j=1}^{P}\mathrm{I}\left(f(x_i) < f(x_j)\right)}{\sum_{h=1}^{M}\mathrm{I}\left(f(x_h) < f(x_i)\right)}$$

- Sigmoid

$$\mathrm{I}\left(f(x_i) < f(x_j)\right) \approx \frac{1}{1 + e^{\alpha\left(f(x_j) - f(x_i)\right)}} = \sigma\left(f(x_j) - f(x_i)\right)$$

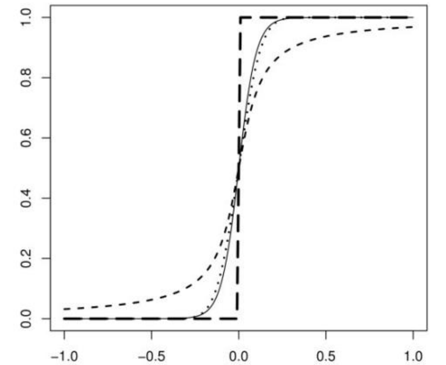
with α a smoothing parameter

Smooth approximation of AP

- Average precision

$$AP = \frac{1}{P} \sum_{i=1}^{P} \frac{\sum_{j=1}^{P} \mathrm{I}\left(f(x_i) < f\left(x_j\right)\right)}{\sum_{h=1}^{M} \mathrm{I}\left(f(x_h) < f(x_i)\right)}$$

- Sigmoid



$$\mathrm{I}\left(f(x_i) < f\left(x_j\right)\right) \approx \frac{1}{1 + e^{\alpha\left(f(x_j) - f(x_i)\right)}} = \sigma\left(f\left(x_j\right) - f(x_i)\right)$$

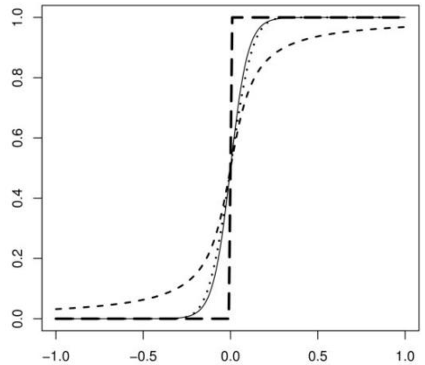
with α a smoothing parameter

$$1 - \widehat{AP}_{sig} = \frac{1}{P} \sum_{i=1}^{P} \frac{\sum_{j=1}^{N} \sigma\left(f\left(x_j\right) - f(x_i)\right)}{\sum_{h=1}^{M} \sigma\left(f(x_h) - f(x_i)\right)}$$

Smooth approximation of AP

- Average precision

$$AP = \frac{1}{P}\sum_{i=1}^{P}\frac{\sum_{j=1}^{P} \mathrm{I}\left(f(x_i) < f(x_j)\right)}{\sum_{h=1}^{M} \mathrm{I}\left(f(x_h) < f(x_i)\right)}$$



- Sigmoid

$$\mathrm{I}\left(f(x_i) < f(x_j)\right) \approx \frac{1}{1 + e^{\alpha\left(f(x_j) - f(x_i)\right)}} = \sigma\left(f(x_j) - f(x_i)\right)$$

with α a smoothing parameter

$$1 - \widehat{AP}_{sig} = \frac{1}{P}\sum_{i=1}^{P}\frac{\sum_{j=1}^{N} \sigma\left(f(x_j) - f(x_i)\right)}{\sum_{h=1}^{M} \sigma\left(f(x_h) - f(x_i)\right)}$$

Complexity : $O(P \times N)$

Smooth approximation of AP

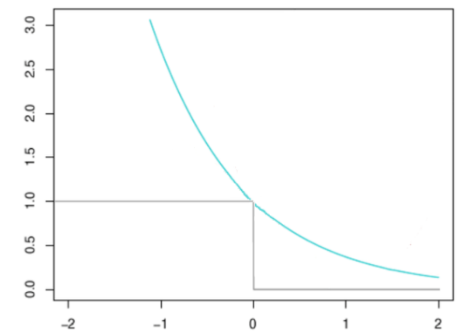
- Average precision

$$AP = \frac{1}{P}\sum_{i=1}^{P}\frac{\sum_{j=1}^{P} \mathrm{I}\left(f(x_i) < f\left(x_j\right)\right)}{\sum_{h=1}^{M} \mathrm{I}\left(f(x_h) < f(x_i)\right)}$$

Smooth approximation of AP

- Average precision

$$AP = \frac{1}{P}\sum_{i=1}^{P}\frac{\sum_{j=1}^{P}\mathrm{I}\left(f(x_i) < f(x_j)\right)}{\sum_{h=1}^{M}\mathrm{I}\left(f(x_h) < f(x_i)\right)}$$

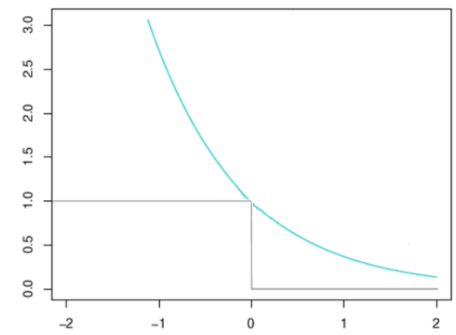


- Exponential

Smooth approximation of AP

- Average precision

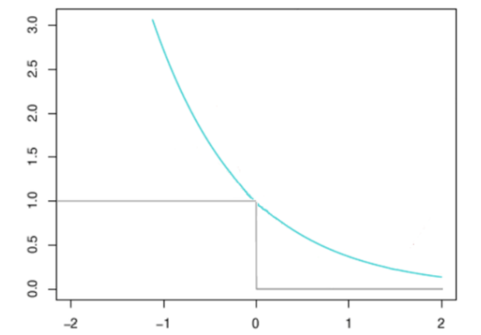
$$AP = \frac{1}{P}\sum_{i=1}^{P} \frac{\sum_{j=1}^{P} \mathrm{I}\left(f(x_i) < f(x_j)\right)}{\sum_{h=1}^{M} \mathrm{I}\left(f(x_h) < f(x_i)\right)}$$



- Exponential

$$\mathrm{I}\left(f(x_i) < f(x_j)\right) \le e^{f(x_j) - f(x_i)}$$

Smooth approximation of AP



- Average precision

$$AP = \frac{1}{P}\sum_{i=1}^{P}\frac{\sum_{j=1}^{P}\mathrm{I}\left(f(x_i) < f(x_j)\right)}{\sum_{h=1}^{M}\mathrm{I}\left(f(x_h) < f(x_i)\right)}$$

- Exponential

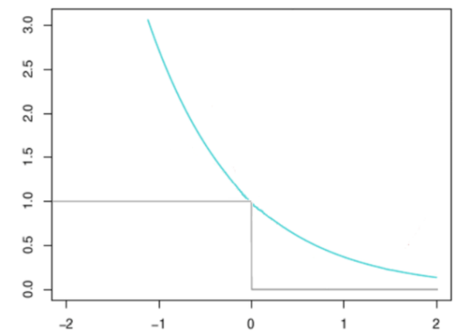
$$\mathrm{I}\left(f(x_i) < f(x_j)\right) \leq e^{f(x_j)-f(x_i)}$$

$$1 - \widehat{AP}_{exp} = \frac{\sum_{n=1}^{N} e^{f(x_n)}}{\sum_{h=1}^{M} e^{f(x_h)}}$$

Smooth approximation of AP

- Average precision

$$AP = \frac{1}{P}\sum_{i=1}^{P}\frac{\sum_{j=1}^{P} \mathrm{I}\left(f(x_i) < f\left(x_j\right)\right)}{\sum_{h=1}^{M} \mathrm{I}\left(f(x_h) < f(x_i)\right)}$$



- Exponential

$$\mathrm{I}\left(f(x_i) < f\left(x_j\right)\right) \leq e^{f\left(x_j\right) - f(x_i)}$$

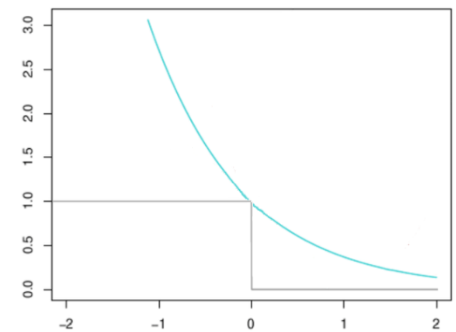
$$1 - \widehat{AP}_{exp} = \frac{\sum_{n=1}^{N} e^{f(x_n)}}{\sum_{h=1}^{M} e^{f(x_h)}}$$

Pros: $O(P + N)$

Smooth approximation of AP

- Average precision

$$AP = \frac{1}{P}\sum_{i=1}^{P}\frac{\sum_{j=1}^{P}\mathrm{I}\left(f(x_i) < f(x_j)\right)}{\sum_{h=1}^{M}\mathrm{I}\left(f(x_h) < f(x_i)\right)}$$



- Exponential

$$\mathrm{I}\left(f(x_i) < f(x_j)\right) \leq e^{f(x_j) - f(x_i)}$$

$$1 - \widehat{AP}_{exp} = \frac{\sum_{n=1}^{N} e^{f(x_n)}}{\sum_{h=1}^{M} e^{f(x_h)}}$$

Pros: $O(P + N)$

Cons: Gradient explosion

Stochastic gradient boosting

Stochastic gradient boosting

Why ?

- Optimizing in function space instead of parameter space
- Adaptive algorithm
- SGB prevents the gradients from exploding

Stochastic gradient boosting

Why ?

- Optimizing in function space instead of parameter space
- Adaptive algorithm
- SGB prevents the gradients from exploding

Weak learners h are combined linearly:

$$f_t(x) = f_{t-1}(x) + \alpha_t h_t(x)$$

Stochastic gradient boosting

At iteration $t-1$ we find the residuals for all $\{x_i\}_{i=1}^{M}$

$$g_t(x_i) = \frac{\partial L(y_i, f_{t-1}(x_i))}{\partial f_{t-1}(x_i)}$$

Stochastic gradient boosting

At iteration $t-1$ we find the residuals for all $\{x_i\}_{i=1}^{M}$

$$g_t(x_i) = \frac{\partial L(y_i, f_{t-1}(x_i))}{\partial f_{t-1}(x_i)}$$

We find a model h_t with its corresponding weight such that:

$$h_t = argmin_h \sum_{i=1}^{M} -g(x_i)h(x_i)$$

$$\alpha_t = argmin_\alpha \sum_{i=1}^{M} L(y_i, f_{t-1}(x_i) + \alpha h_t(x_i))$$

Experiments

- Datasets

	#examples	Positives ratio	#Features
Pima	767	34%	8
Breast cancer	286	30%	9
HIV	3,272	13.3%	8
Heart cleveland (4 vs all)	303	4.3%	13
w8a	64000	3%	300
Fraud	2,000,000	0.2%	40

Experiments

- Datasets

	#examples	Positives ratio	#Features
Pima	767	34%	8
Breast cancer	286	30%	9
HIV	3,272	13.3%	8
Heart cleveland (4 vs all)	303	4.3%	13
w8a	64000	3%	300
Fraud	2,000,000	0.2%	40

- Evaluation metrics:
  - AUCROC
  - Average precision

Experiments

- Datasets

	#examples	Positives ratio	#Features
Pima	767	34%	8
Breast cancer	286	30%	9
HIV	3,272	13.3%	8
Heart cleveland (4 vs all)	303	4.3%	13
w8a	64000	3%	300
Fraud	2,000,000	0.2%	40

- Evaluation metrics:
  - AUCROC
  - Average precision
  - Pos@Top, the number of positives before the first negative appears

Experiments

- Datasets

	#examples	Positives ratio	#Features
Pima	767	34%	8
Breast cancer	286	30%	9
HIV	3,272	13.3%	8
Heart cleveland (4 vs all)	303	4.3%	13
w8a	64000	3%	300
Fraud	2,000,000	0.2%	40

- Evaluation metrics:
  - AUCROC
  - Average precision
  - Pos@Top, the number of positives before the first negative appears
  - P@k, with k equal to the number of positive examples

Experiments

- Datasets

	#examples	Positives ratio	#Features
Pima	767	34%	8
Breast cancer	286	30%	9
HIV	3, 272	13.3%	8
Heart cleveland (4 vs all)	303	4.3%	13
w8a	64000	3%	300
Fraud	2, 000, 000	0.2%	40

- Evaluation metrics:
  - AUCROC
  - Average precision
  - Pos@Top, the number of positives before the first negative appears
  - P@k, with k equal to the number of positive examples

- Models for comparison

Experiments

- Datasets

	#examples	Positives ratio	#Features
Pima	767	34%	8
Breast cancer	286	30%	9
HIV	3,272	13.3%	8
Heart cleveland (4 vs all)	303	4.3%	13
w8a	64000	3%	300
Fraud	2,000,000	0.2%	40

- Evaluation metrics:
  - AUCROC
  - Average precision
  - Pos@Top, the number of positives before the first negative appears
  - P@k, with k equal to the number of positive examples
- Models for comparison
  - GB-logistic (Friedman 2001)

Experiments

- Datasets

	#examples	Positives ratio	#Features
Pima	767	34%	8
Breast cancer	286	30%	9
HIV	3,272	13.3%	8
Heart cleveland (4 vs all)	303	4.3%	13
w8a	64000	3%	300
Fraud	2,000,000	0.2%	40

- Evaluation metrics:
  - AUCROC
  - Average precision
  - Pos@Top, the number of positives before the first negative appears
  - P@k, with k equal to the number of positive examples
- Models for comparison
  - GB-logistic (Friedman 2001)
  - Rankboost (Freund et al 2003)

Experiments

- Datasets

	#examples	Positives ratio	#Features
Pima	767	34%	8
Breast cancer	286	30%	9
HIV	3,272	13.3%	8
Heart cleveland (4 vs all)	303	4.3%	13
w8a	64000	3%	300
Fraud	2,000,000	0.2%	40

- Evaluation metrics:
  - AUCROC
  - Average precision
  - Pos@Top, the number of positives before the first negative appears
  - P@k, with k equal to the number of positive examples

- Models for comparison
  - GB-logistic (Friedman 2001)
  - Rankboost (Freund et al 2003)
  - LambdaMART-AP (Burges 2010)

Experiments

- Datasets

	#examples	Positives ratio	#Features
Pima	767	34%	8
Breast cancer	286	30%	9
HIV	3, 272	13.3%	8
Heart cleveland (4 vs all)	303	4.3%	13
w8a	64000	3%	300
Fraud	2, 000, 000	0.2%	40

- Evaluation metrics:
  - AUCROC
  - Average precision
  - Pos@Top, the number of positives before the first negative appears
  - P@k, with k equal to the number of positive examples

- Models for comparison
  - GB-logistic (Friedman 2001)
  - Rankboost (Freund et al 2003)
  - LambdaMART-AP (Burges 2010)
  - SGBAP

Results

Dataset	Algorithm	$AUCROC$	AP	$Pos@Top$	$P@k$
Pima	GB-Logistic		71.3%	3.9%	66.1%
	RankBoost		72.8%	6.2%	65.9%
	LambdaMART-AP		73.4%	4.1%	65.6%
	SGBAP		71.2%	5.8%	64.6%
Breast cancer	GB-Logistic		50.9%	9.3%	44.6%
	RankBoost		48.4%	4.6%	46.3%
	LambdaMART-AP		52.8%	8.6%	51.9%
	SGBAP		56%	10.2%	49.8%
HIV	GB-Logistic		55.6%	3%	53.9%
	RankBoost		54.6%	4%	53.1%
	LambdaMART-AP		42.9%	0.8%	48.7%
	SGBAP		57.4%	5.4%	54.5%
Heart cleveland	GB-Logistic		16.4%	1.3%	10%
	Rankboost		17.4%	1.5%	9.7%
	LambdaMART-AP		18.1%	3.8%	13.3%
	SGBAP		21.9%	4.8%	20.2%
w8a	GB-Logistic		73.8%	5.3%	70.9%
	RankBoost		76.5%	3.9%	72.7%
	LambdaMART-AP		–	–	–
	SGBAP		83.5%	17.8%	79.7%
Fraud	GB-Logistic		14.7%	0.09%	24.1%
	RankBoost		15.6%	0.05%	24.5%
	LambdaMART-AP		–	–	–
	SGBAP		17.5%	0.6%	32%

Results

Dataset	Algorithm	$AUCROC$	AP	$Pos@Top$	$P@k$
Pima	GB-Logistic	82.8%	71.3%	3.9%	66.1%
	RankBoost	83.5%	72.8%	6.2%	65.9%
	LambdaMART-AP	81.8%	73.4%	4.1%	65.6%
	SGBAP	82.8%	71.2%	5.8%	64.6%
Breast cancer	GB-Logistic	68.2%	50.9%	9.3%	44.6%
	RankBoost	64.9%	48.4%	4.6%	46.3%
	LambdaMART-AP	67.3%	52.8%	8.6%	51.9%
	SGBAP	71.2%	56%	10.2%	49.8%
HIV	GB-Logistic	85.9%	55.6%	3%	53.9%
	RankBoost	85.9%	54.6%	4%	53.1%
	LambdaMART-AP	82.2%	42.9%	0.8%	48.7%
	SGBAP	86.6%	57.4%	5.4%	54.5%
Heart cleveland	GB-Logistic	75.4%	16.4%	1.3%	10%
	Rankboost	81.1%	17.4%	1.5%	9.7%
	LambdaMART-AP	72.7%	18.1%	3.8%	13.3%
	SGBAP	77.9%	21.9%	4.8%	20.2%
w8a	GB-Logistic	95.4%	73.8%	5.3%	70.9%
	RankBoost	97.1%	76.5%	3.9%	72.7%
	LambdaMART-AP	–	–	–	–
	SGBAP	97%	83.5%	17.8%	79.7%
Fraud	GB-Logistic	88.1%	14.7%	0.09%	24.1%
	RankBoost	88.3%	15.6%	0.05%	24.5%
	LambdaMART-AP	–	–	–	–
	SGBAP	68.8%	17.5%	0.6%	32%

Conclusion

- We proposed a learning to rank approach for anomaly detection problems

Conclusion

- We proposed a learning to rank approach for anomaly detection problems
- One of our approximations is linear and can scale to big datasets

Conclusion

- We proposed a learning to rank approach for anomaly detection problems
- One of our approximations is linear and can scale to big datasets
- As the data are unbalanced, experiments show that our method performs better in the top rank

Perspectives

- Automatic decision threshold based on expert criteria

Perspectives

- Automatic decision threshold based on expert criteria
- Adaptation to online learning

Perspectives

- Automatic decision threshold based on expert criteria
- Adaptation to online learning
- Open the proprietary dataset with an international data science competition

Thank you for your attention