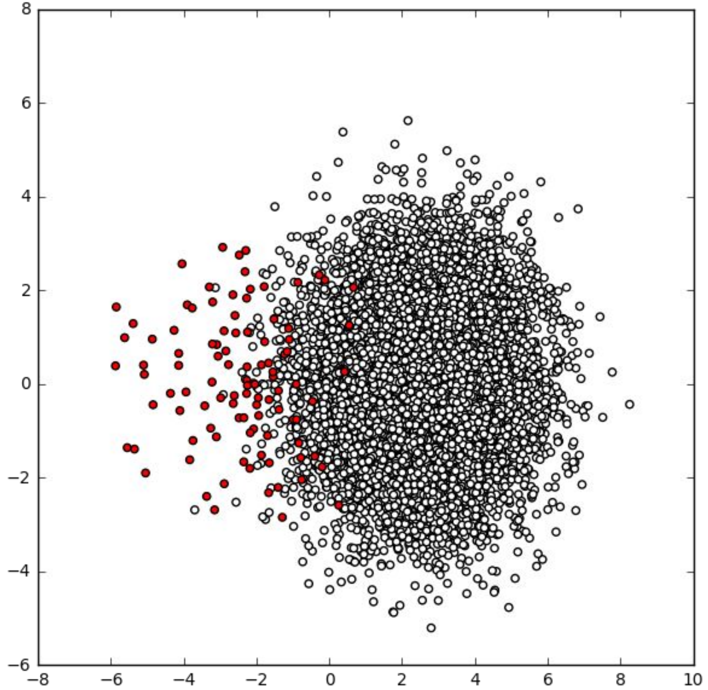
Profit Curves
& Imbalanced Classes

17-01-DS-SEA
Galvanize, Seattle
Jfomhover

Credits: drawing on work from Ryan Henning, Ivan Corneillet, Darren Reger...



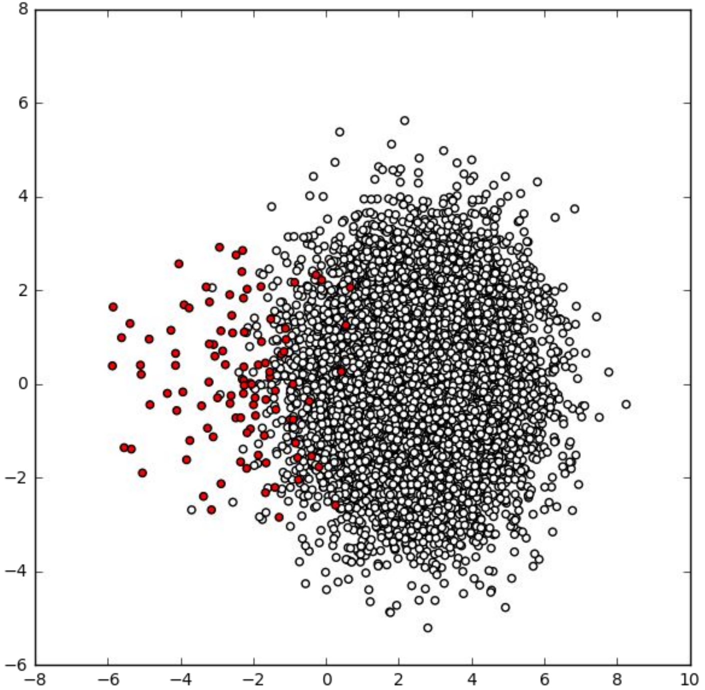
Profit Curves
& Imbalanced Classes

17-01-DS-SEA
Galvanize, Seattle
jfomhover

OBJECTIVES

- **Discuss** and give examples of the issues with imbalanced classes.
- **Explain** and **implement** the profit curve method.
- **Explain** cost sensitive learning and how it deals with imbalanced classes.
- **Define**, give examples and relate sampling methods.

Imbalanced Classes: failure analysis

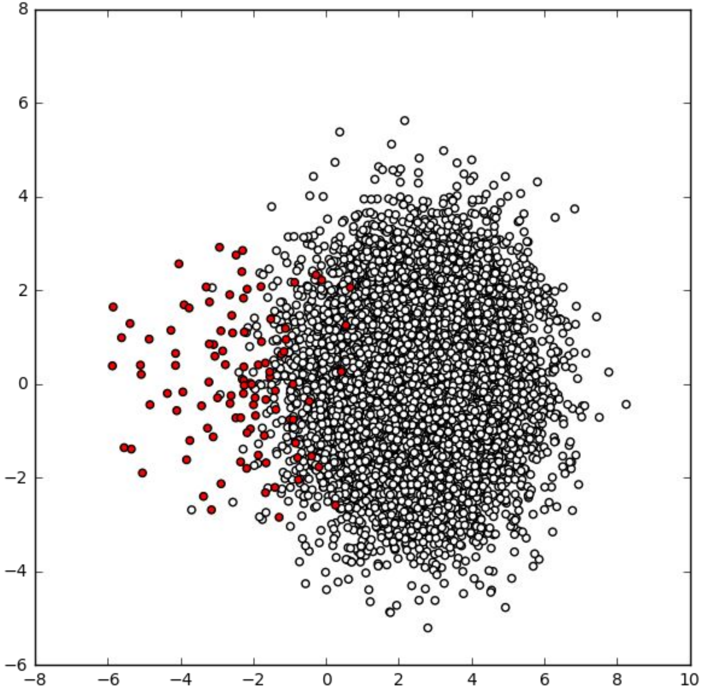


Pb : what could it change during LEARNING ?

Pb: what could it change during EVALUATION ?

Example : 100 pos, 10000 neg

Imbalanced Classes: failure analysis



Example : 100 pos, 10000 neg

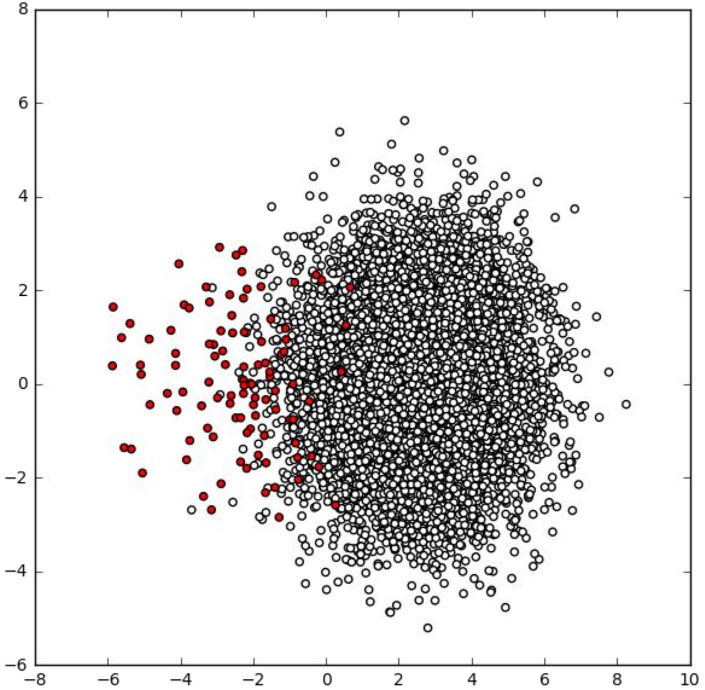
Pb : what <u>could</u> it change during LEARNING ?

Sol: cost-sensitive learning, over/under sampling

Pb: what <u>could</u> it change during EVALUATION ?

Sol: cost-benefit matrix

Imbalanced Classes during EVALUATION



Example : 100 pos, 10000 neg

I can design a classifier with 99% accuracy !

Accuracy-driven models will over-predict the majority class.

QUESTION: how would you pick your favorite matrix ?

A	Pred: pos	Pred: neg
Actual: pos	12	15
Actual: neg	8	965

B	Pred: pos	Pred: neg
Actual: pos	0	0
Actual: neg	20	980

C	Pred: pos	Pred: neg
Actual: pos	15	115
Actual: neg	5	865

D	Pred: pos	Pred: neg
Actual: pos	18	250
Actual: neg	2	730

QUESTION: how would you pick your favorite matrix ?

A	Pred: pos	Pred: neg
Actual: pos	12	15
Actual: neg	8	965

B	Pred: pos	Pred: neg
Actual: pos	0	[illegible]
[illegible]	[illegible]	980

C	[illegible]	[illegible]
[illegible]	15	115
Actual: neg	5	865

D	Pred: pos	Pred: neg
Actual: pos	18	250
Actual: neg	2	730

DEFINE A BUSINESS PROBLEM

QUESTION: how would you pick your favorite matrix ?

A	Pred: pos	Pred: neg
Actual: pos	12	15
[illegible]	[illegible]	[illegible]

B	Pred: pos	Pred: neg
Actual: pos	[illegible]	[illegible]
[illegible]	[illegible]	980

C	[illegible]	[illegible]
[illegible]	[illegible]	115
Actual: neg	5	865

[illegible]	[illegible]	[illegible]
Actual: pos	18	[illegible]
Actual: neg	2	730

FORMALIZE COSTS AND BENEFITS

DEFINE A BUS[illegible]SS PROBLEM

Confusion Matrix

	Pred: Y	Pred: N
Actual: y	TP	FN
Actual: n	FP	TN

Confusion Matrix

P = TP+FN = count of actual y
N = FP+TN = count of actual n

VALUES ARE COUNTS

Probability Matrix

	Pred: Y	**Pred: N**
Actual: y	TP	FN
Actual: n	FP	TN

Confusion Matrix

P = TP+FN = count of actual y
N = FP+TN = count of actual n

VALUES ARE COUNTS

	Pred: Y	**Pred: N**
Actual: y	p(Y,y)	p(N,y)
Actual: n	p(Y,n)	p(N,n)

Probability Matrix

$$p(Y,y) = TP / (P + N)$$
$$p(Y,n) = FP / (P + N)$$
$$p(N,y) = FN / (P + N)$$
$$p(N,n) = TN / (P + N)$$

VALUES ARE PROBAS

Cost-Benefit Matrix

	Pred: Y	Pred: N
Actual: y	TP	FN
Actual: n	FP	TN

Confusion Matrix

P = TP+FN = count of actual y
N = FP+TN = count of actual n

VALUES ARE COUNTS

	Pred: Y	Pred: N
Actual: y	p(Y,y)	p(N,y)
Actual: n	p(Y,n)	p(N,n)

Probability Matrix

p(Y,y) = TP / (P + N)
p(Y,n) = FP / (P + N)
p(N,y) = FN / (P + N)
p(N,n) = TN / (P + N)

VALUES ARE PROBAS

	Pred: Y	Pred: N
Actual: y	b(Y,y)	c(N,y)
Actual: n	c(Y,n)	b(N,n)

Cost-Benefit Matrix

VALUES ARE $$$!

Computing the Expected Profit

	Pred: Y	Pred: N
Actual: y	TP	FN
Actual: n	FP	TN

	Pred: Y	Pred: N
Actual: y	p(Y,y)	p(N,y)
Actual: n	p(Y,n)	p(N,n)

	Pred: Y	Pred: N
Actual: y	b(Y,y)	c(N,y)
Actual: n	c(Y,n)	b(N,n)

$$E[Profit] = p(Y,y).\,b(Y,y) + p(Y,n).\,c(Y,n) + p(N,y).\,c(N,y) + p(N,n).\,b(N,n)$$

$$= p(Y \mid y).\,p(y).\,b(Y,p) + p(Y \mid n).\,p(n).\,c(Y,n) + p(N \mid y).\,p(y).\,c(N,y) + p(N \mid n).\,p(n).\,b(N,n)$$

$$= p(y).\,[p(Y \mid y).\,b(Y,p) + p(N \mid y).\,c(N,y)] + p(n)\,[p(Y \mid n).\,c(Y,n) + p(N \mid n).\,b(N,n)]$$

Cost-Benefit Matrix (example 1)

Prompt: You are building a model to predict if credit card charges are fraudulent.

- If we predict a fraudulent charge, we'll call the customer to confirm.
- If you miss a fraudulent charge, it on average costs $100
- Calling someone to confirm if their charge was real costs on average $4

Question: What is an appropriate cost benefit matrix?

A	**Predicted: fraud**	**Predicted: not fraud**
Actual: fraud	96	-100
Actual: not fraud	-4	0

B	**Predicted: fraud**	**Predicted: not fraud**
Actual: fraud	-4	-100
Actual: not fraud	-4	0

C	**Predicted: fraud**	**Predicted: not fraud**
Actual: fraud	96	0
Actual: not fraud	-4	0

Cost-Benefit Matrix (example 2)

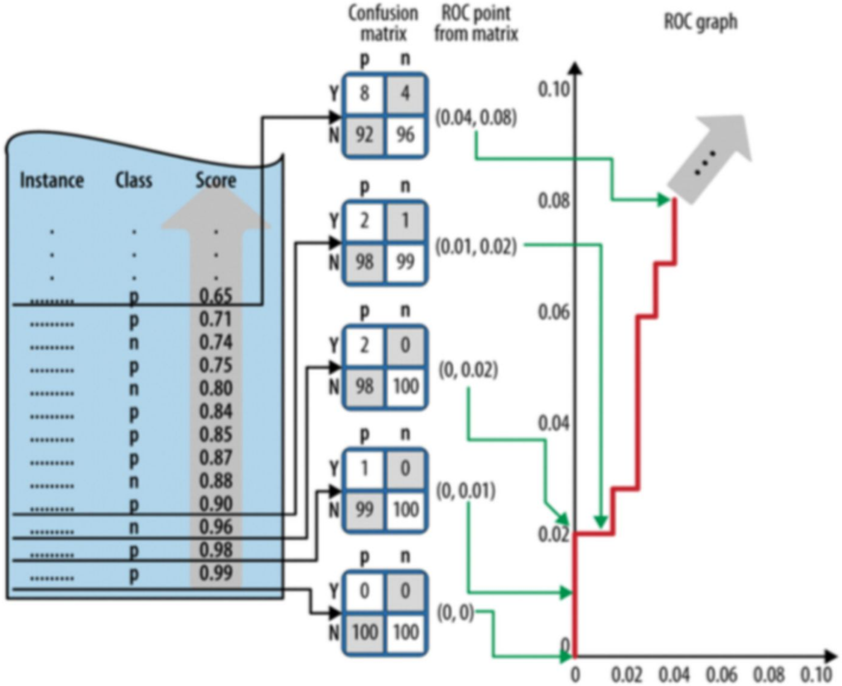
You are building a model to **predict if customers will churn** from your online clothing store.
You'll use your model **to send a promotional email** to users you think are going to churn.

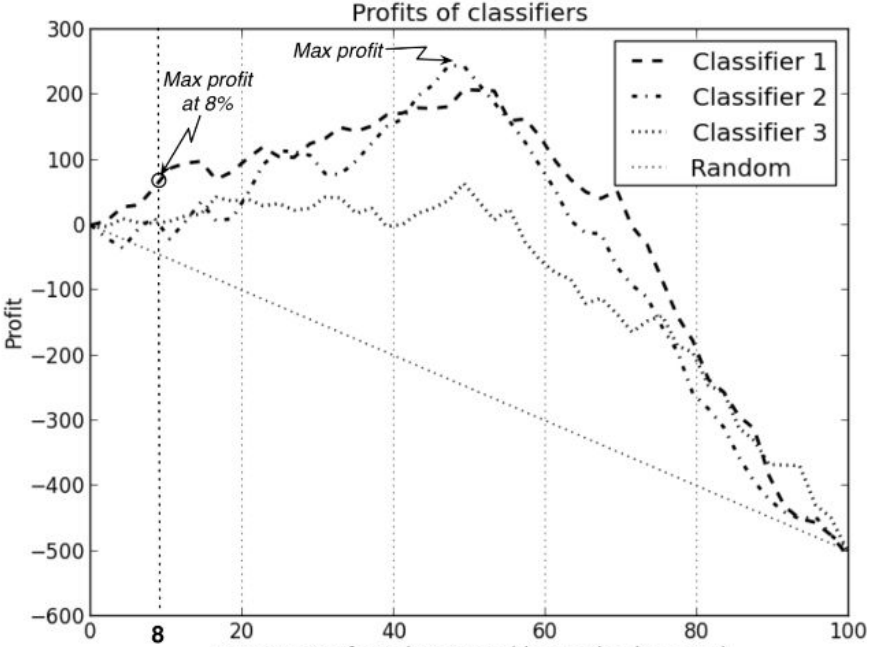
You'd like to use a cost benefit matrix so you can build **profit curves to determine the optimal model**.

- Customers on average spend **$200/month**.
  Your profit is **10%** of this revenue.
- A promotional email costs on average **$2/customer** and prevents **50%** of users from churning for **6 months**.
- When the promotional email is sent to users who were not going to churn, it annoys **5%** of them and causes them to churn **2 months** earlier than they otherwise would have.

	Predicted: churn	**Predicted: not churn**
Actual: churn	?	?
Actual: Not churn	?	?

From Thresholding to Profit Curves



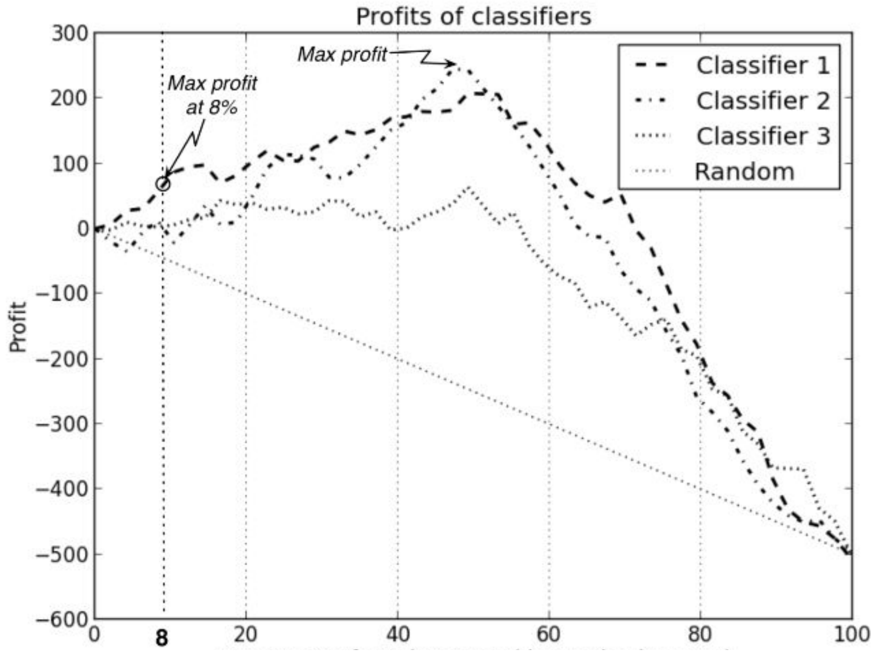


Percent of test instances (decreasing by score)

Compare classifiers on one given cost benefit matrix

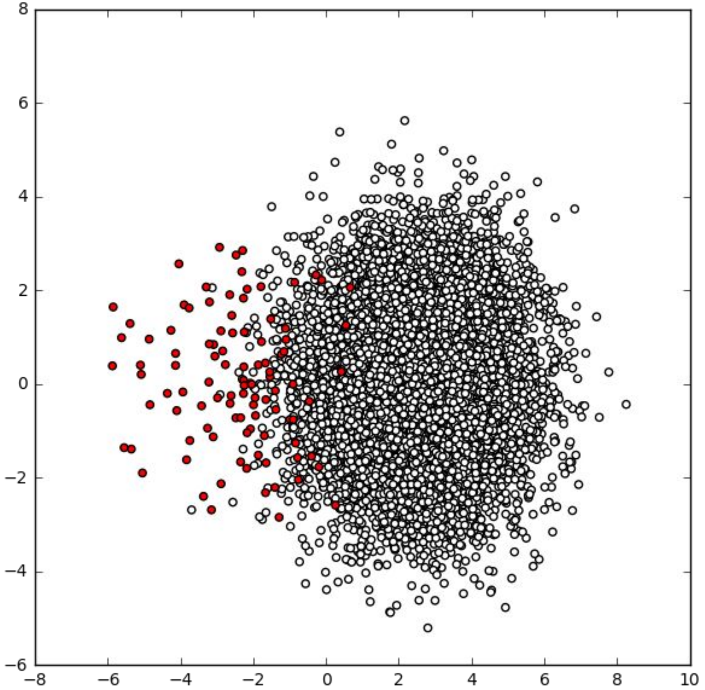
Profit Curve:

- Same idea as ROC curve but with expected profit
- For each threshold, compute the expected profit



Percent of test instances (decreasing by score)

Imbalanced Classes: failure analysis



Example : 100 pos, 10000 neg

Pb : what <u>could</u> it change during LEARNING ?

Sol: cost-sensitive learning, over/under sampling

Pb: what <u>could</u> it change during EVALUATION ?

Sol: cost-benefit matrix

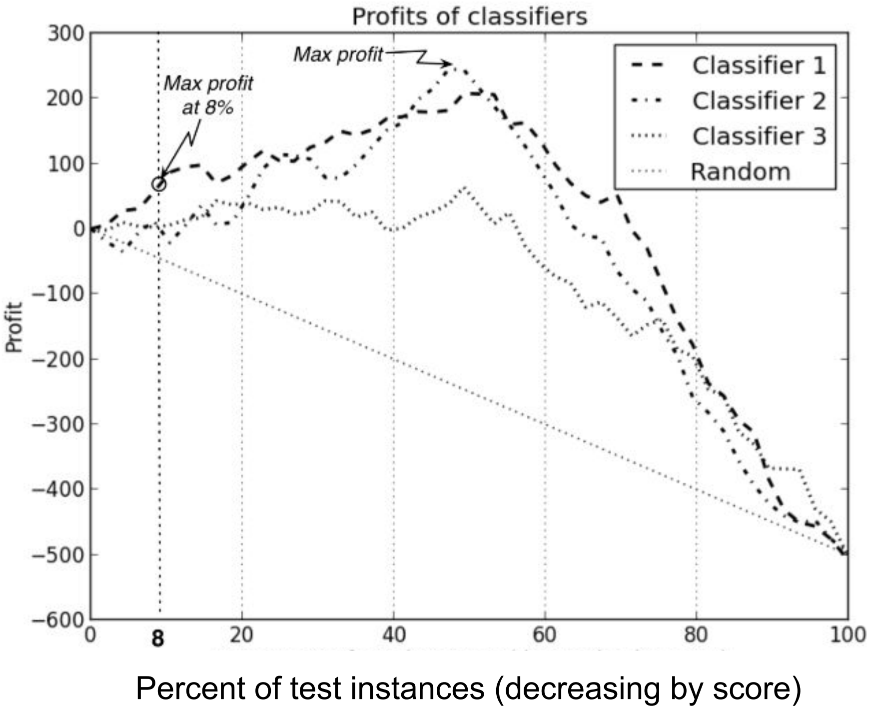
Compare classifiers on one given cost benefit matrix

Profit Curve:

- Same idea as ROC curve but with expected profit
- For each threshold, compute the expected profit

Cost-sensitive learning:

- Select threshold with highest expected profit.



Introducing cost-benefit in the objective function

- Models with explicit objective function can be modified to incorporate classification cost.

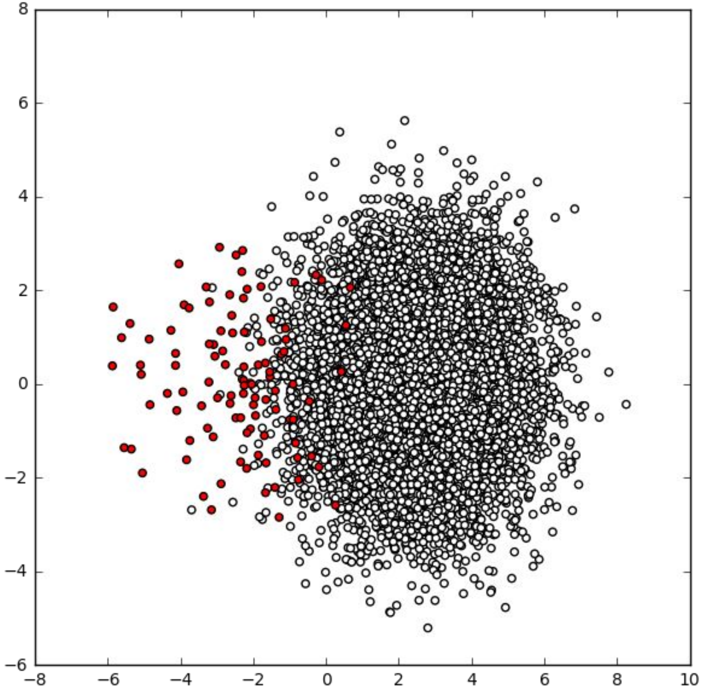
$$\ln p(\vec{y}|X;\theta) = \sum_{i=1}^{n} \left(y_i \ln h_\theta(x_i) + (1-y_i)\ln(1-h_\theta(x_i))\right)$$

- e.g. logistic regression
  This will affect optimization.
  - cost-sensitive logistic regression may not be convex anymore !

$$J^c(\theta) = \frac{1}{N}\sum_{i=1}^{N}\Big(y_i(h_\theta(X_i)C_{TP_i} + (1-h_\theta(X_i))C_{FN_i}) + (1-y_i)(h_\theta(X_i)C_{FP_i} + (1-h_\theta(X_i))C_{TN_i})\Big).$$

- Not all models have a cost-sensitive implementation.

Imbalanced Classes: failure analysis



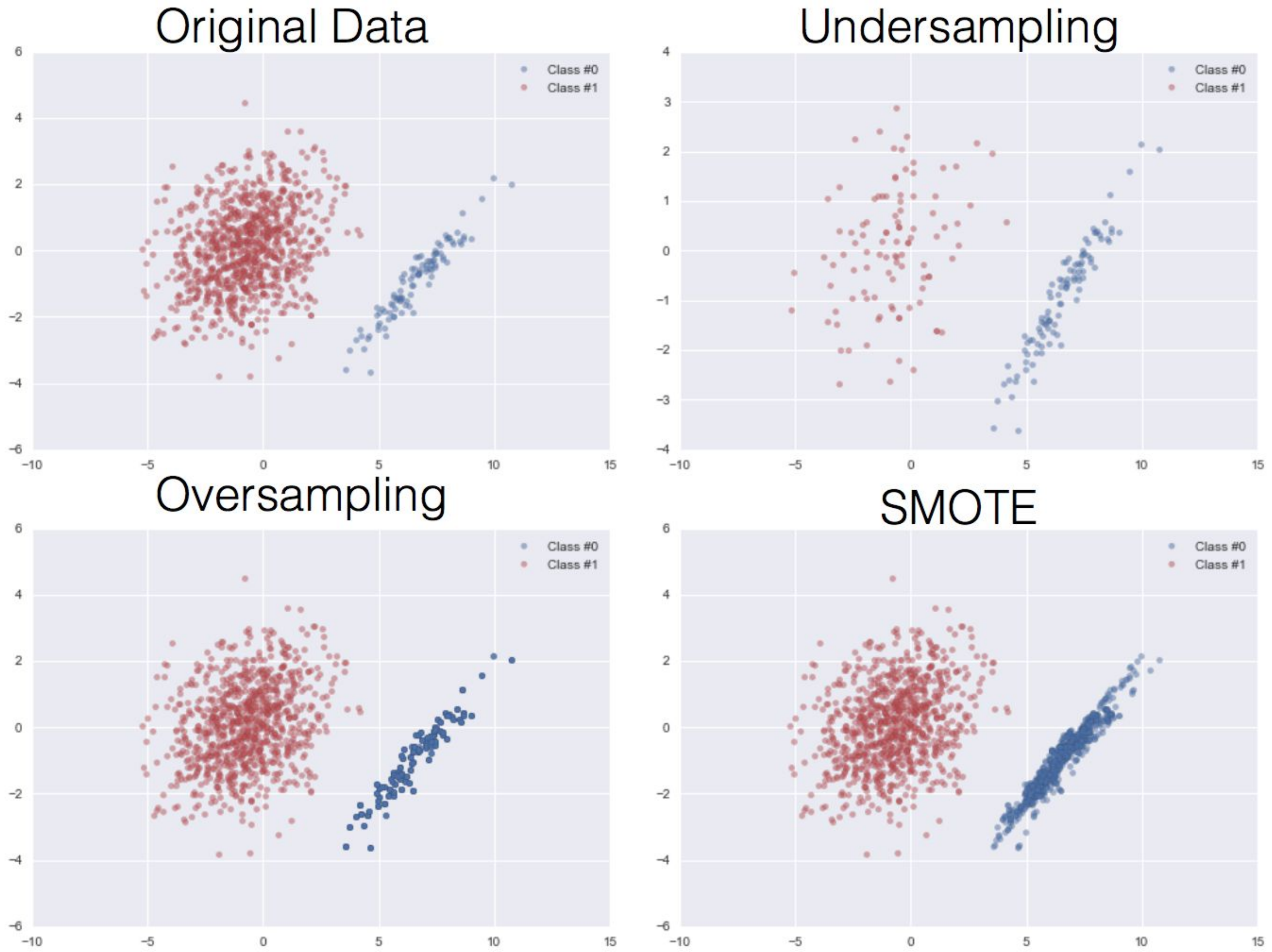
Example : 100 pos, 10000 neg

Pb : what could it change during LEARNING ?

Sol: cost-sensitive learning, over/under sampling

Pb: what could it change during EVALUATION ?

Sol: cost-benefit matrix



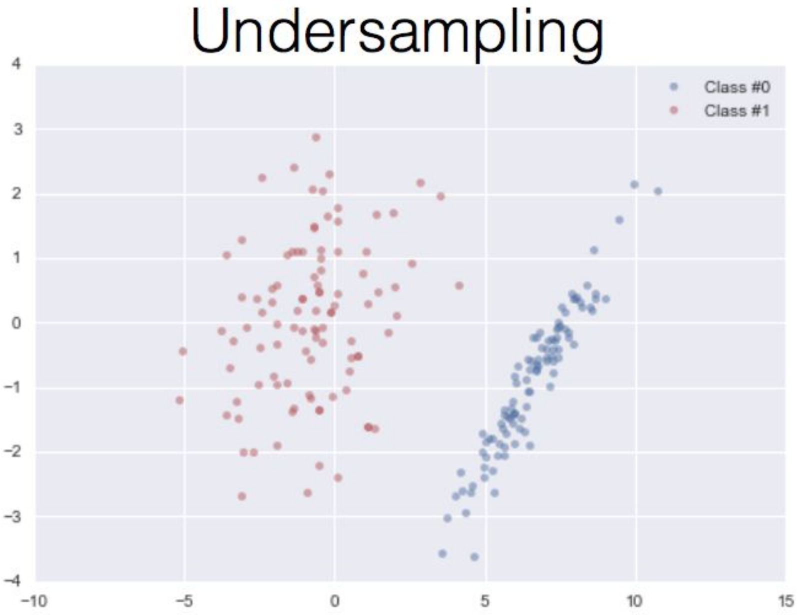
Original Data
Undersampling
Oversampling
SMOTE
Class #0
Class #1

Undersampling

Undersampling randomly discards majority class observations to balance training sample.

PRO: Reduces runtime on very large datasets.

CON: Discards potentially important observations.

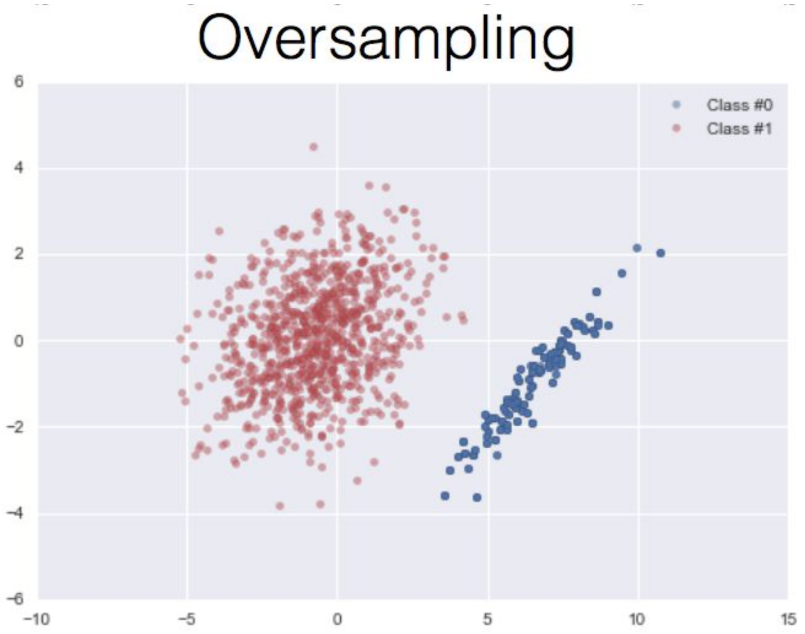


Oversampling

Oversampling replicates observations from minority class to balance training sample.

PRO: Doesn't discard information.

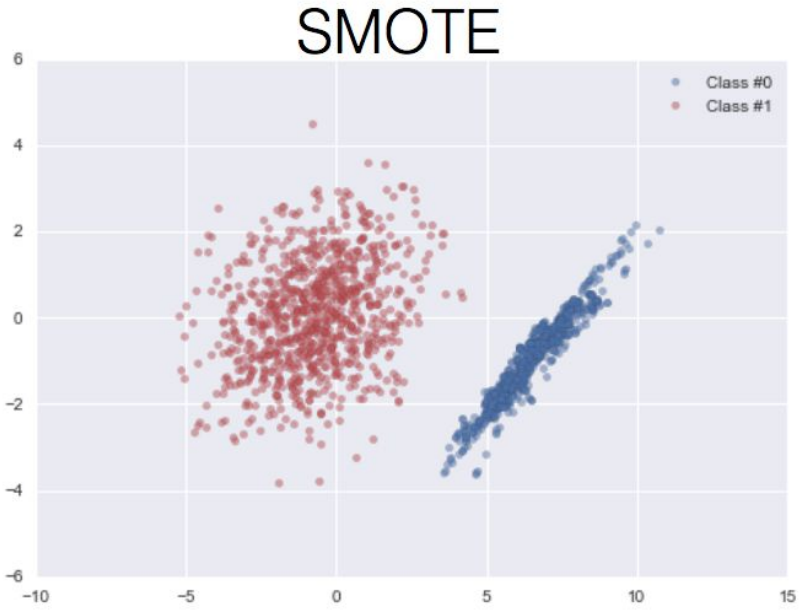
CON: Likely to overfit.



SMOTE - Synthetic Minority Oversampling TEchnique

Generates new observations from minority class.

For each minority class observation and for each feature, randomly generate between it and one of its k-nearest neighbors.



SMOTE pseudocode

```
synthetic_observations = []
while len(synthetic_observations) + len(minority_observations) < target:
    obs = random.choice(minority_observations):
    neighbor = random.choice(kNN(obs, k)) # randomly selected neighbor
    new_observation = {}
    for feature in obs:
        weight = random() # random float between 0 and 1
        new_feature_value = weight*obs[feature] \
                            + (1-weight)*neighbor[feature]
        new_observation[feature] = new_feature_value
    synthetic_observations.append(new_observation)
```

QUESTIONS

Consider these 3 scenarios:

1) You are building a model to determine if credit card charges are **fraudulent**.
   You have the data for **10,000** credit card charges and **100** of them are fraudulent.

2) You are building to model to determine if a picture is of **a dog or a cat**.
   You have **40,000** pictures of dogs and **10,000** pictures of cats.

3) You are building a model to **detect spam emails**.
   You have **1,000,000** emails and **25,000** of the emails are spam.

In each of these scenarios,

- What percent of the data points is the minority class?
- What should you do in each of these scenarios?
  Would you use any of SMOTE, undersampling or oversampling?
- What questions might you want to ask about your data to help facilitate determining the answer?

Profit Curves & Imbalanced Classes

17-01-DS-SEA
Galvanize, Seattle
jfomhover

OBJECTIVES

- **Discuss** and give examples of the issues with imbalanced classes.
- **Explain** and **implement** the profit curve method.
- **Explain** cost sensitive learning and how it deals with imbalanced classes.
- **Define**, give examples and relate sampling methods.