PFIセミナー

2012/01/05

大岩 秀和 @kisa12012

自己紹介

@kisa12012

- 大岩 秀和
  - 2010年夏期PFIインターン
    - クラスタリングライブラリ等を改良
- 東京大学情報理工学系研究科修士2年
- 専門は機械学習
  - オンライン学習・確率的最適化等

本日のテーマ

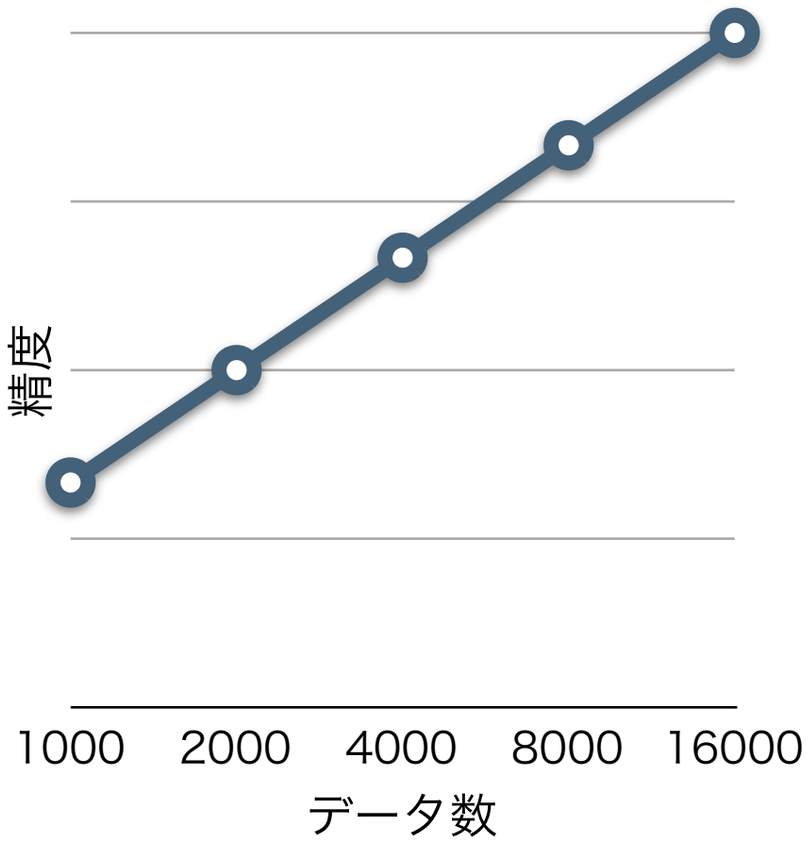
- 能動学習 [Active Learning]
  - 機械学習の一分野
- 教師データの効率のよい作り方の話

埴輪

埴輪

背景

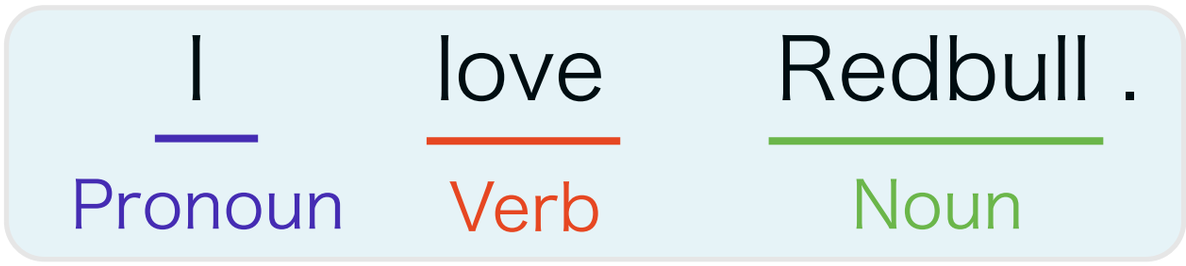
- 学習の精度はデータ量の対数スケールに比例して上昇
  - 精度向上に要求されるデータ量は指数的に増大
    - 構文解析 [Becker+, IJCAI2005]
    - 統計的機械翻訳 [Brant+, EMNLP2007]



大量のデータのラベル付けが
要求される

ラベル付けのコスト

- ラベル付けは（多くのタスクで）高コスト
  - 時間
  - お金（人を雇う必要も）
  - データストレージの管理
    - KDD Tutorial 2011の例：1ヶ月5万ドル20人で100万データ
- ラベル付けが必要な例
  - ニューステキストのカテゴリ分類
  - 画像／映像のタグ付け・セグメンテーション
  - 文章のアノテーション

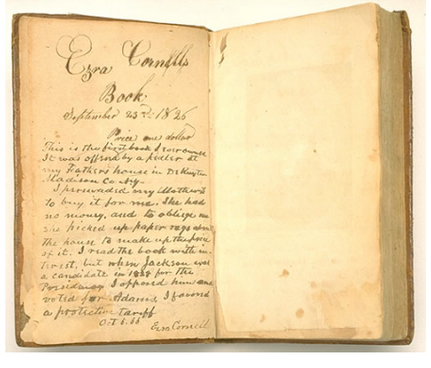


能動学習

- 学習に有用なデータを選択する手法
  - 有用なデータにだけラベル付け
    - 最小の労力で最大の成果を！

通常の教師あり学習

問題集と解答集が
与えられる

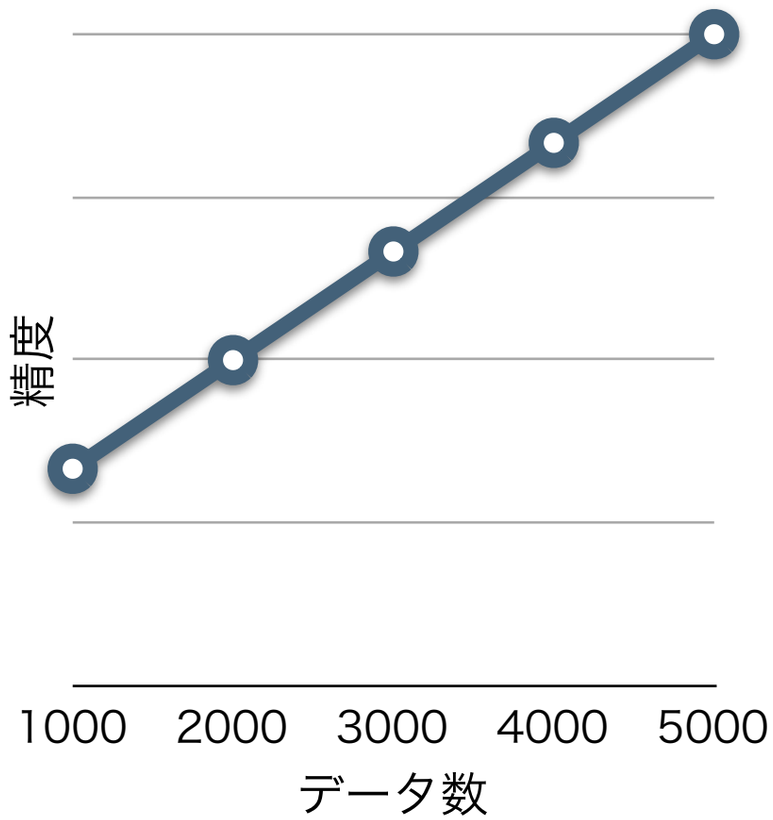


能動学習

先生に分からない
箇所を逐一聞く

能動学習の効果

- 少量のデータで，高速に学習可能
  - データ量に線形に精度が向上するケースも
  - ラベル無しのデータは指数的に必要



ラベル付けのコストが
格段に低下

線形になる例

- (-∞,∞)上のある整数を当てるゲーム
  - 毎ターン，1つの数字を宣言する
  - 宣言した数字が正解ならば終了
  - 間違いの時は，正解より小さいか大きいかが分かる
  - 1次元特徴空間の2値分類問題
- 通常の教師あり学習：適当に数字を宣言
  - エラー率は対数線形に減少
- 能動学習：二分探索
  - エラー率は線形に減少

能動学習が特に有効な例

- ラベル付けに専門家の知識が必要
  - 音声認識 [Zhu+, 2005]
  - 高位のアノテーション（商品名や人名） [Settles, 2008]
  - 製薬 [Warmuth+ 2003]
- ラベル付けに長い時間が必要
  - 薬の効果測定
- タスク毎に粒度が変化する場合
  - 高位のアノテーション（極性判定／避難経路抽出）

能動学習の種類

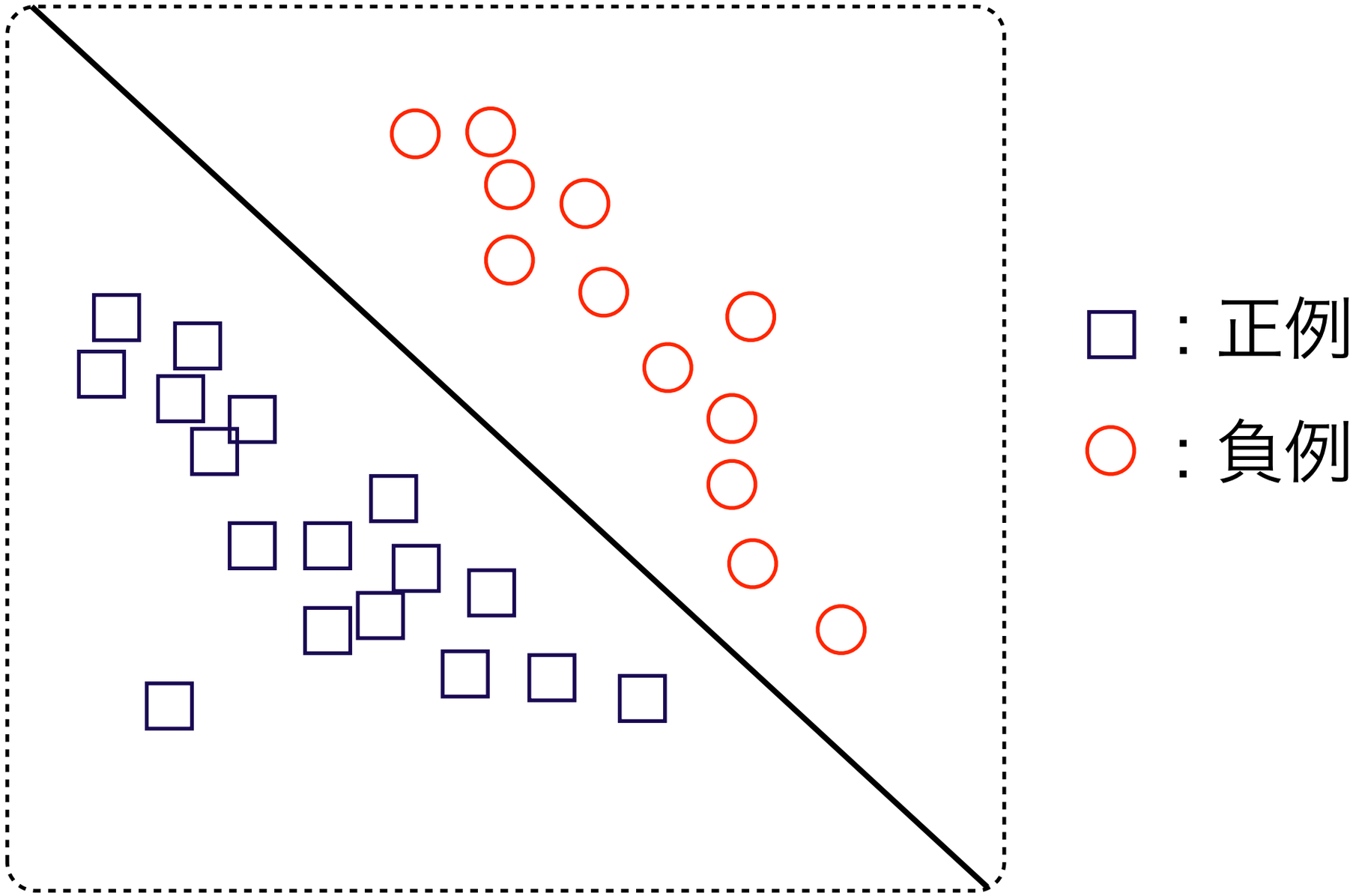
- プールベース能動学習 (Pool-based Active Learning)
  - 最も一般的な手法
- ストリームベース能動学習 (Stream-based Active Learning)
  - 近年注目されている [Beygelzimer+, ICML2009] [Beygelzimer+, NIPS2010]
  - オンライン学習と相性が良い
- クエリ生成型能動学習 (Membership Query Synthesis)
  - 自然言語処理では使われにくい

プールベース能動学習

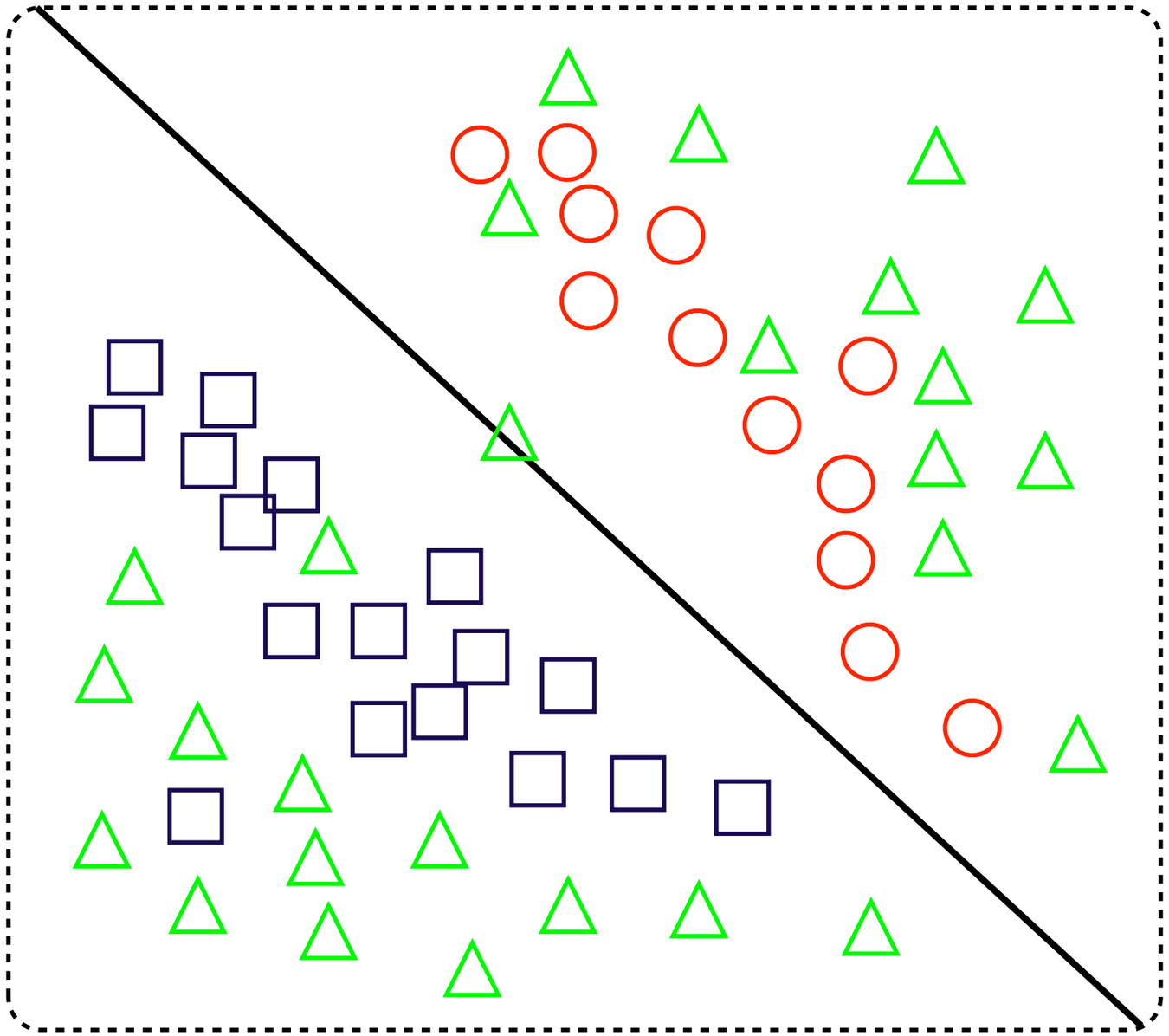
- ラベル無しデータをプールに大量に貯蓄
  - 現在のモデルにおいて，学習に最も有用なデータをプールの中から選択
- ラベル付けしたデータを用いてモデルを更新
- 上の繰り返し

モデル

データプール



学習に有用なラベル無データを選択

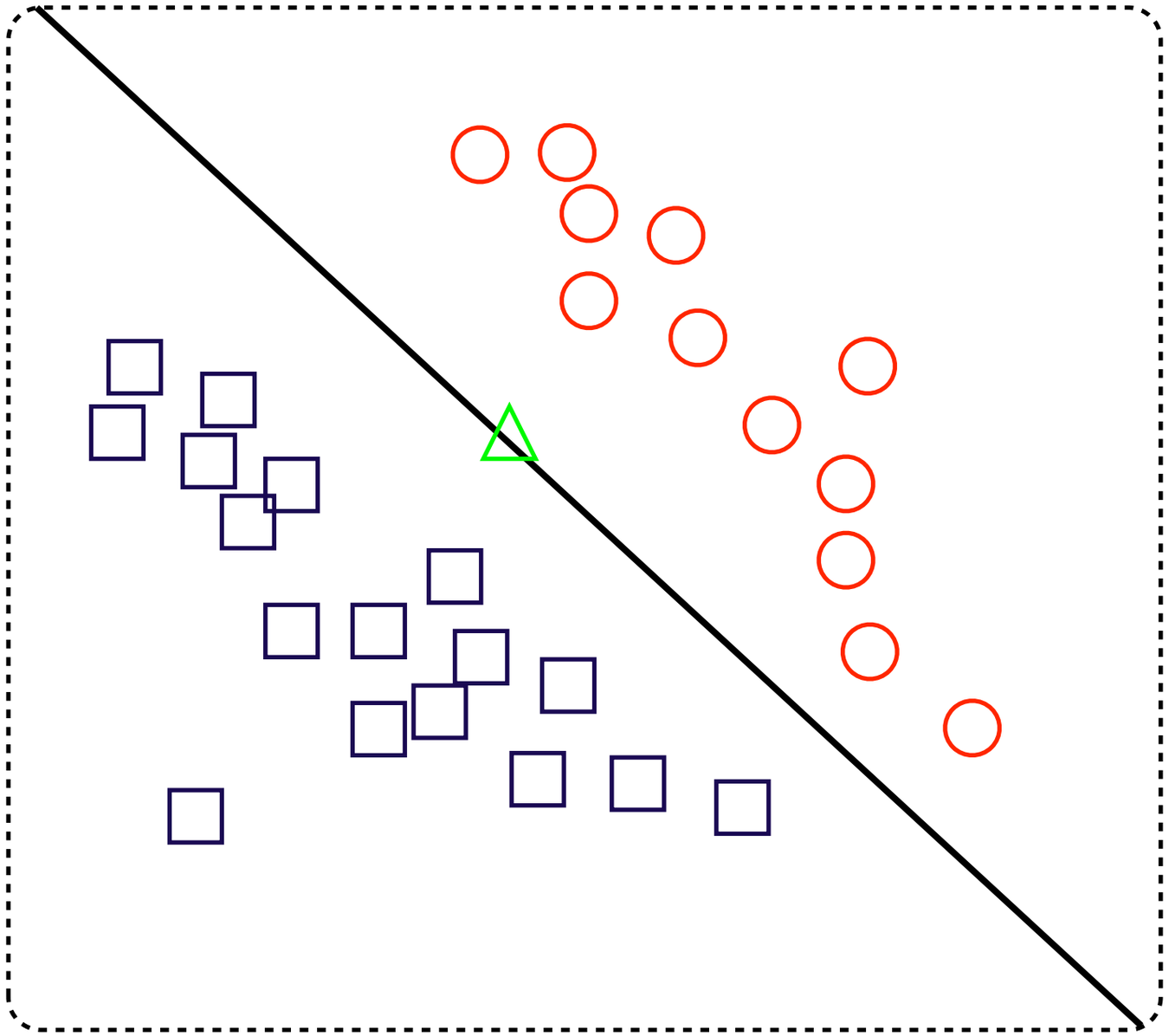


□：正例

○：負例

△：ラベル無

学習に有用なラベル無データを選択

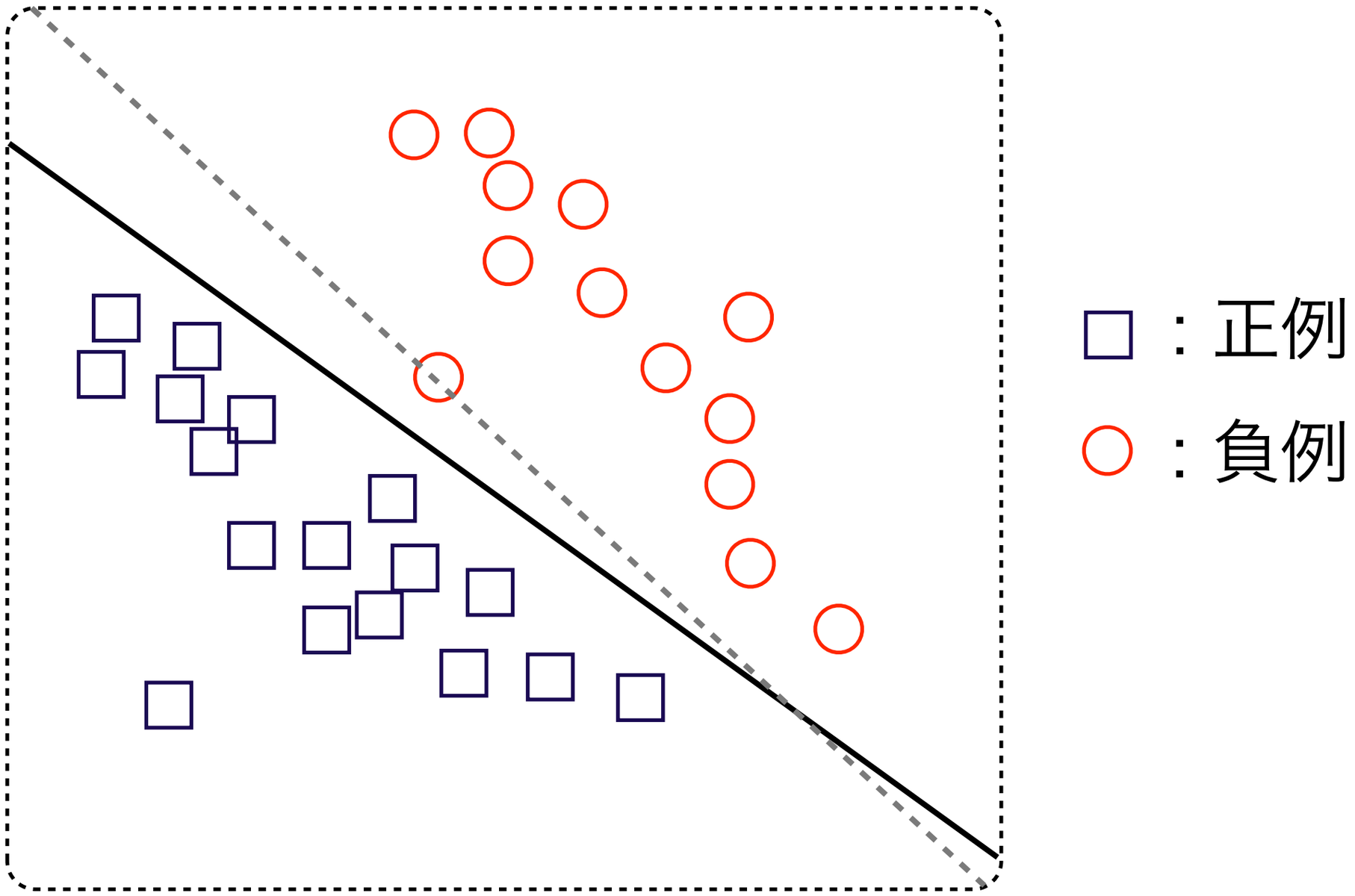


□：正例

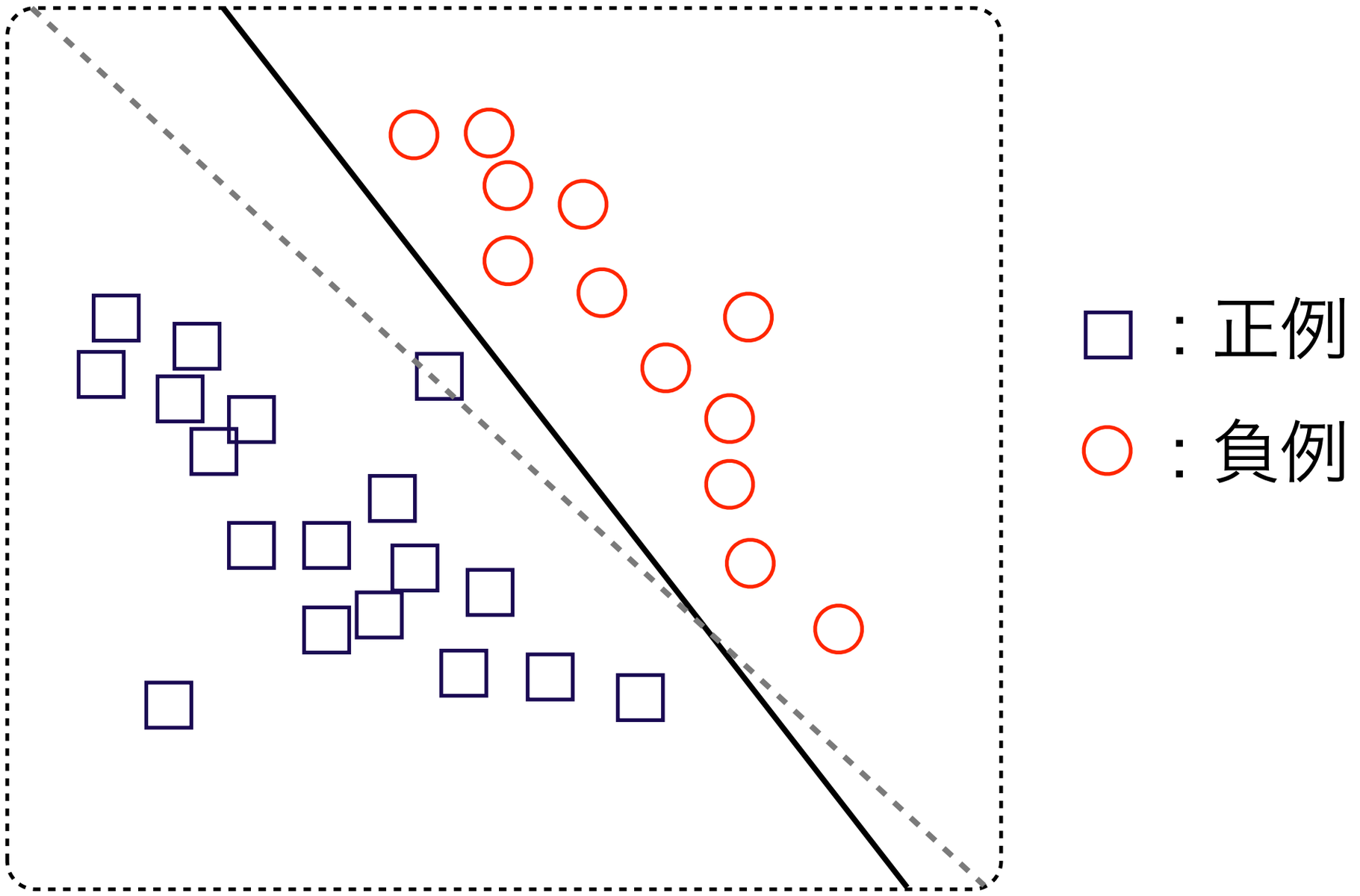
○：負例

△：ラベル無

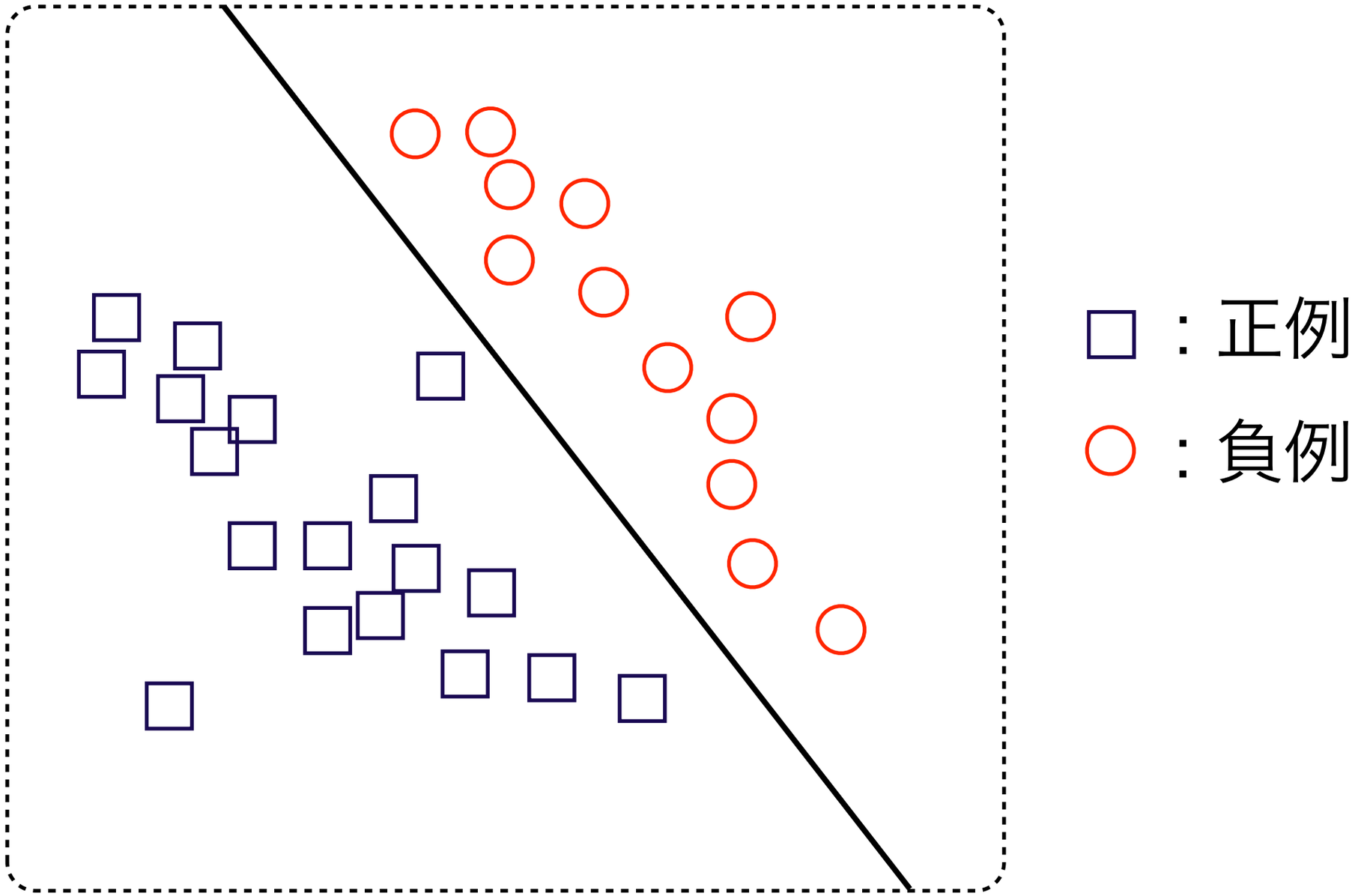
学習に有用なラベル無データを選択



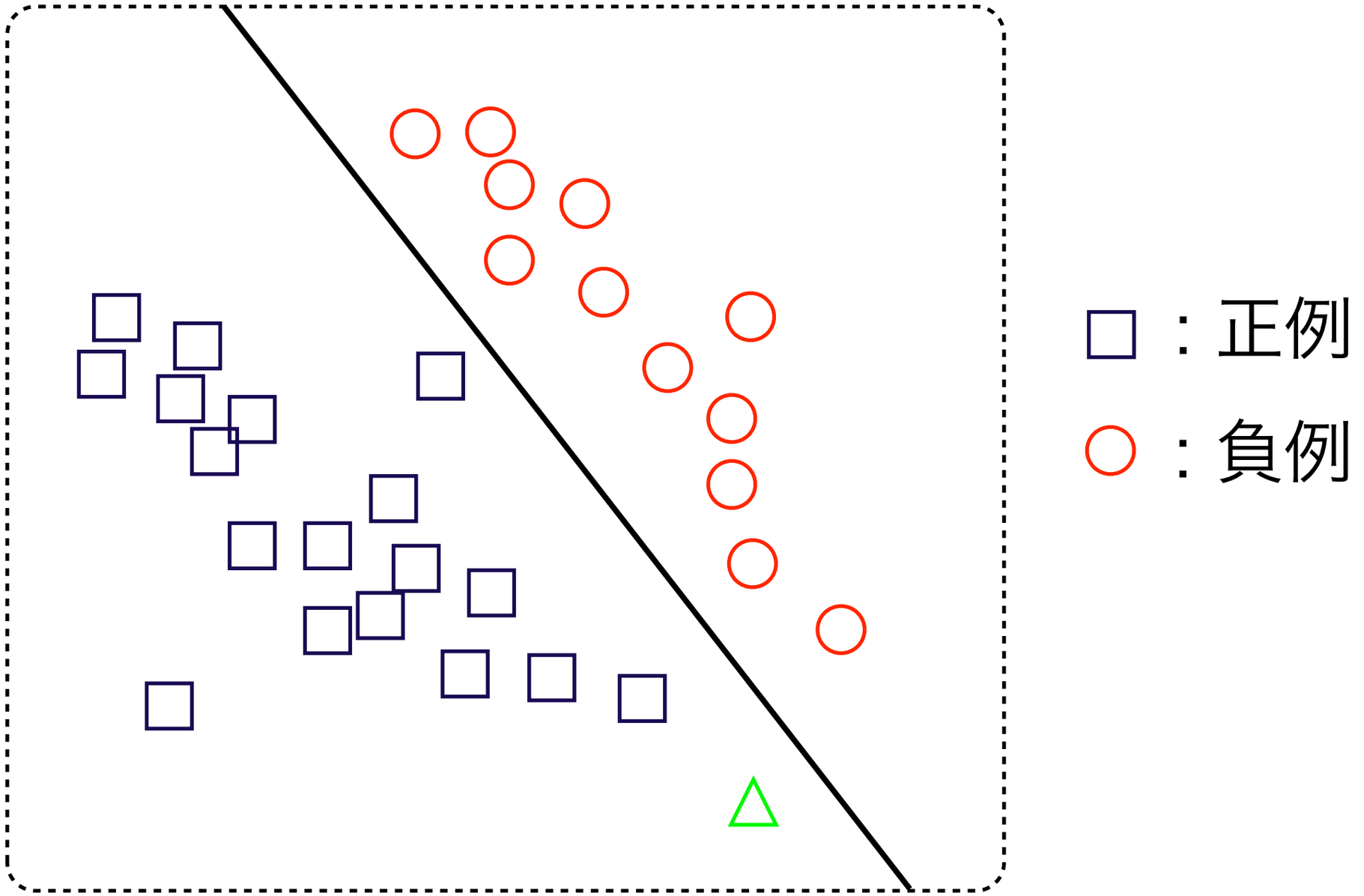
学習に有用なラベル無データを選択



学習に有用なラベル無しデータを選択



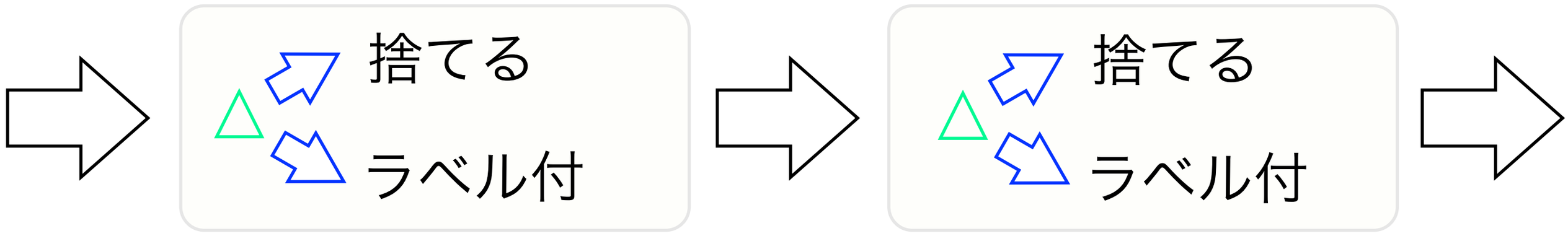
学習に有用なラベル無しデータを選択



学習に有用なラベル無しデータを選択

ストリームベース能動学習

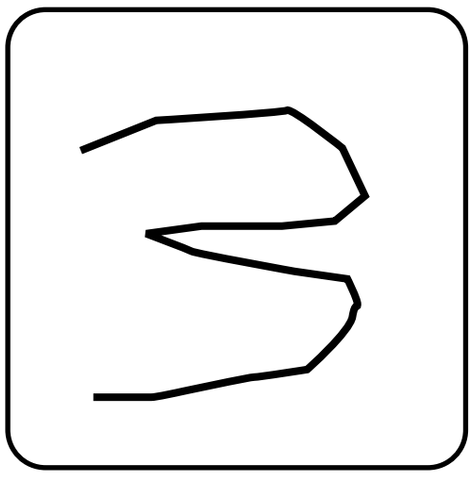
- データが一つ与えられる度，ラベル付するか否かをその場で判断
- ラベル付の有無と無関係にデータは廃棄
- データを貯められない場合やストリーム的にデータを扱いたい場合に有効



クエリ生成型能動学習

- 学習に有用なデータを自分で生成

例：手書き数字認識

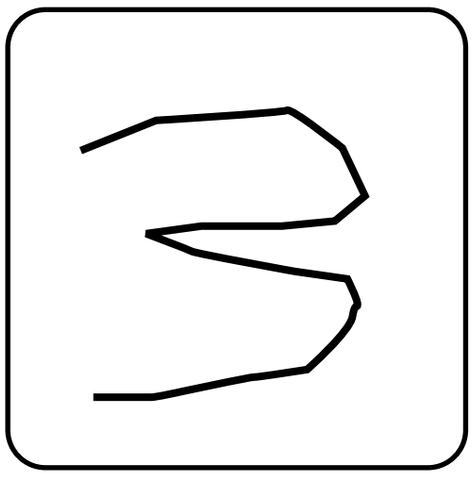


正解：３

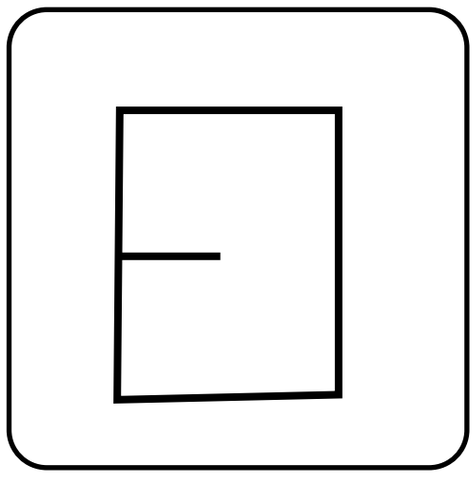
クエリ生成型能動学習

- 学習に有用なデータを自分で生成

例：手書き数字認識



正解：３



正解：？

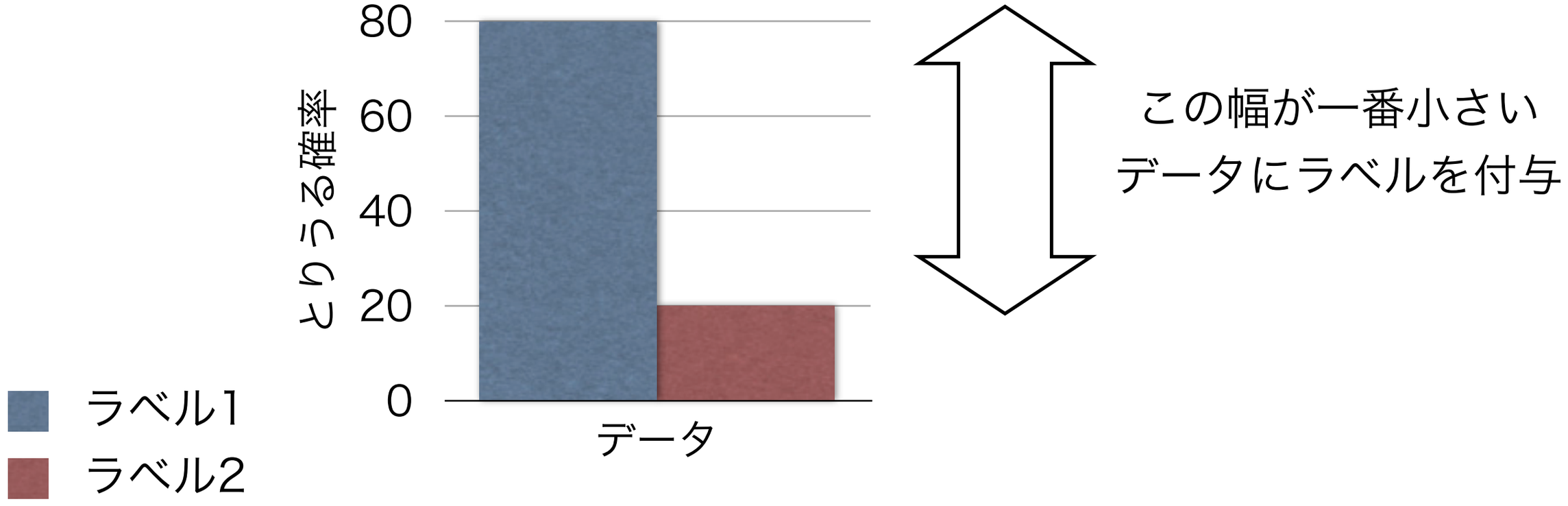
０と８の間のデータを生成

データ選択の基準

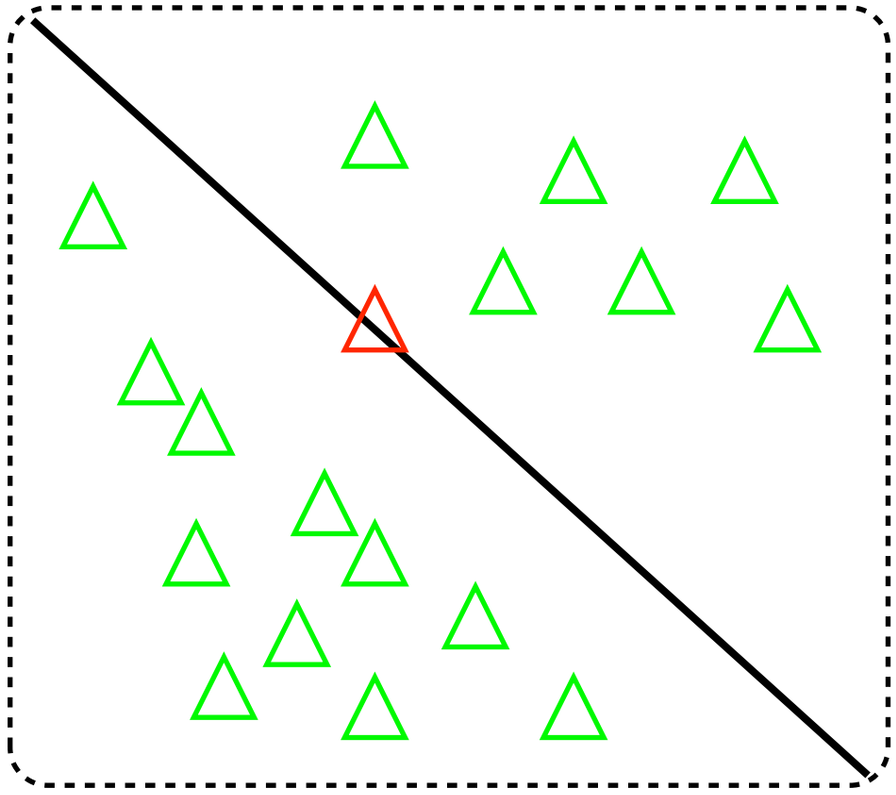
- 非常に沢山の手法が提案されている
  - Query-By-Committee, Expected Model Change, Expected Error Reduction, Variance Reduction, Agnostic Active Learning...
- Uncertainty Samplingを紹介
  - 一番ベーシックな手法
  - 現実的な計算量でデータ選択が可能
  - マージン方式／エントロピー方式
  - 生成モデルが念頭に置かれる $P(y|\mathbf{x})$

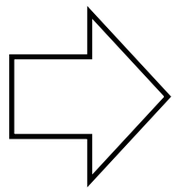
Uncertainty Sampling (Margin) [Scheffer+, CAIDA2001]

- マージンが最も小さいデータを選択



分離平面に一番近いデータを選ぶイメージ





△のデータに
ラベルを付与

式にすると

$$\mathbf{x}_{next} = \underset{\mathbf{x} \in U}{\arg\min}\, P_\theta(\hat{y}_1|\mathbf{x}) - P_\theta(\hat{y}_2|\mathbf{x})$$

$\mathbf{x}_{next}$	次にラベル付けするデータ
U	ラベル無しデータの集合
θ	現在のモデル（学習器）
y	ラベル
$\hat{y}_1$	最も取りうる確率の高いラベル
$\hat{y}_2$	二番目に取りうる確率の高いラベル

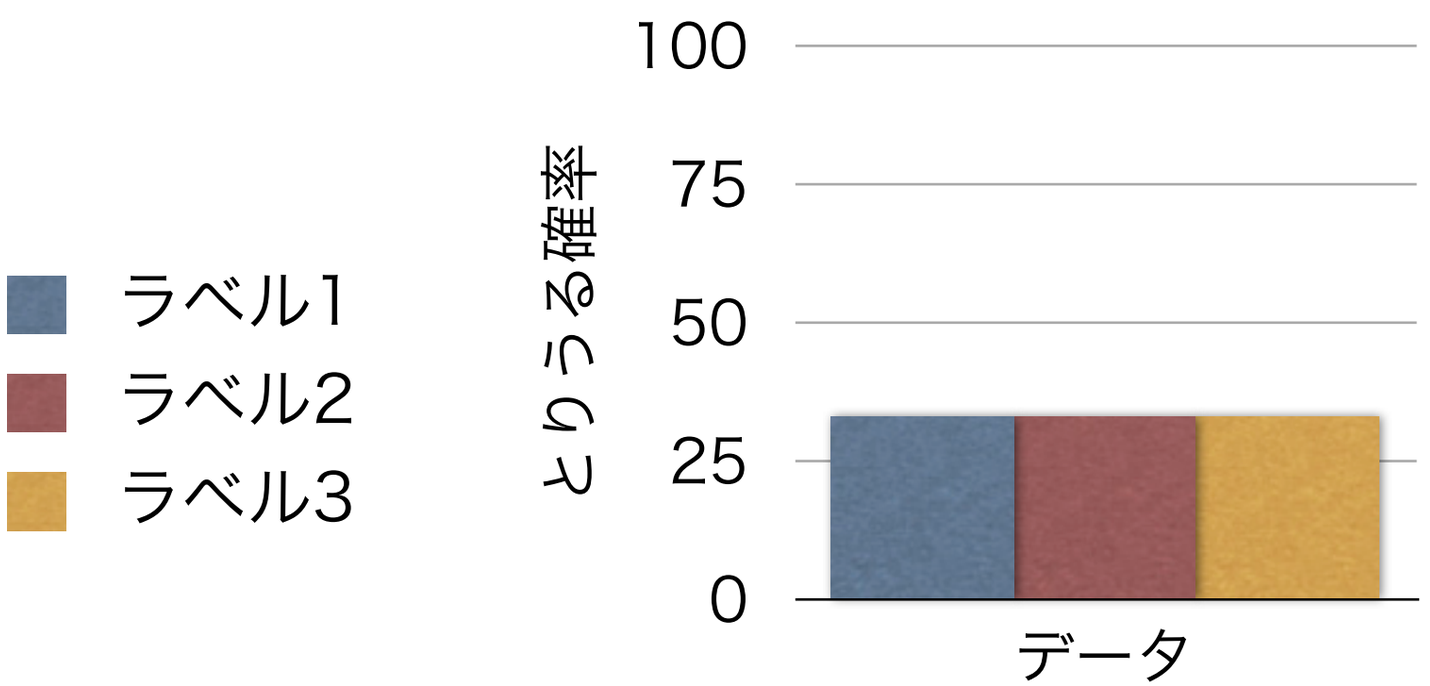
Uncertainty Sampling (Entropy) [Dagan+, ICML95]

- マージン基準は，クラス数が3つ以上ある場合に上位2つのクラスの情報しか使用しない
- 全クラスの情報を使って，最も不確かなデータにラベルを付与したい
  - エントロピー [シャノン情報量]

式にすると

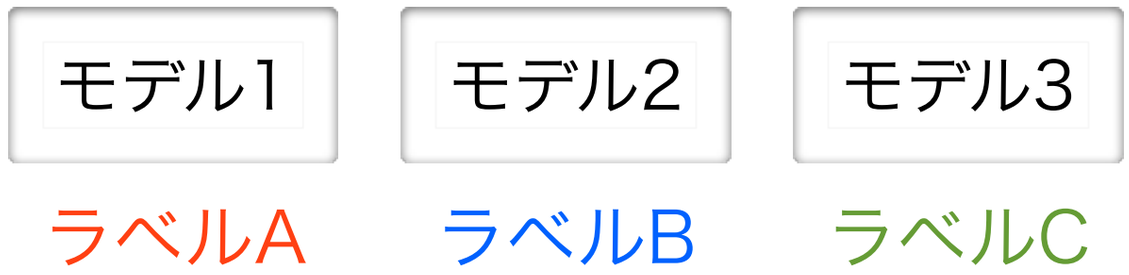
$$\mathbf{x}_{next} = \arg\max_{\mathbf{x}} -\sum_{y} P_\theta(y|\mathbf{x}) \log P_\theta(y|\mathbf{x})$$

エントロピー最大のデータを選択



その他の基準

- Query-By-Committee (disagreement)
  - バージョン空間を上手く等分するデータを選択
  - 複数のモデルを同時に学習させながら，モデル間で予測されるラベルが異なるデータを選択する方法がよく取られる



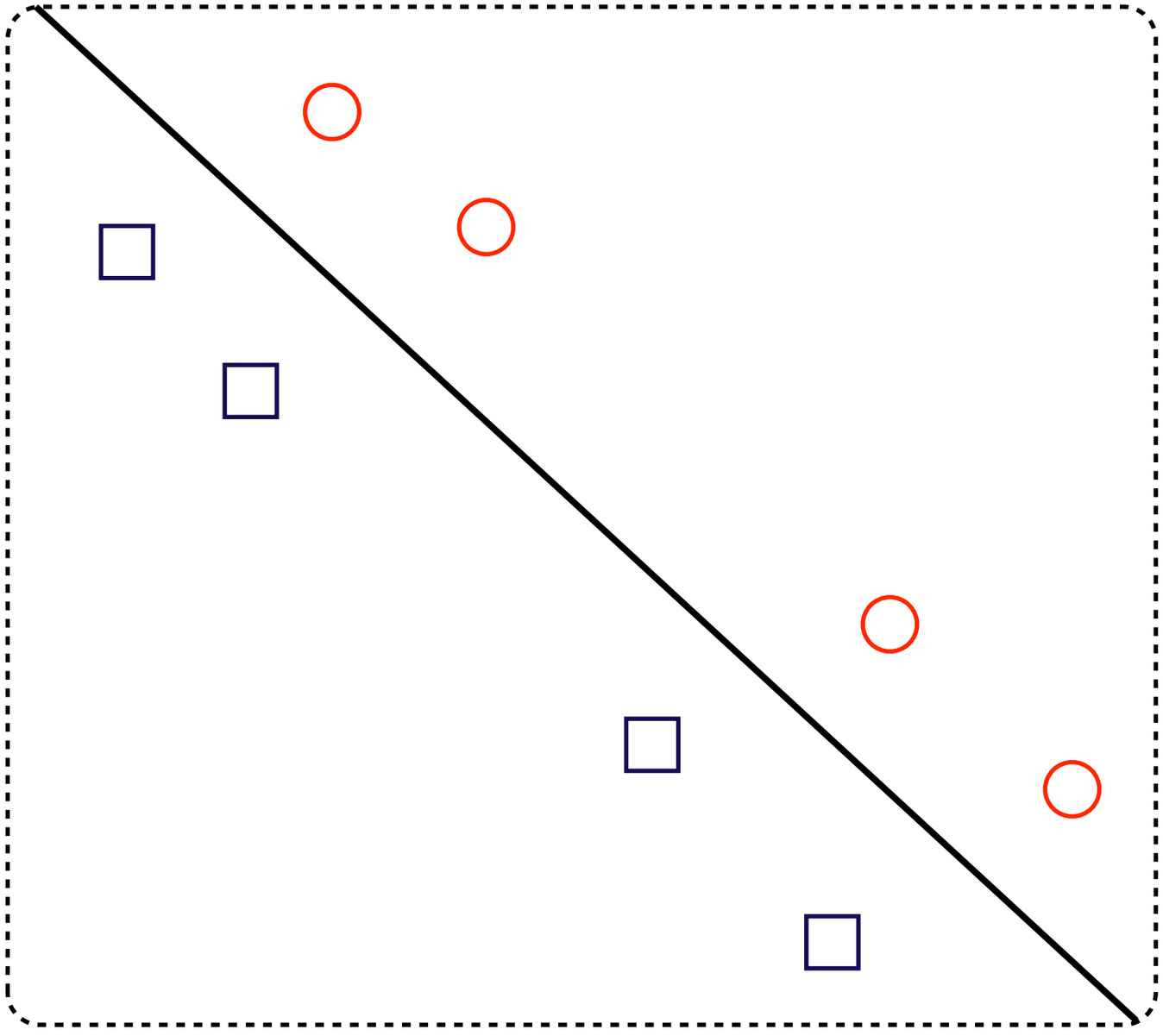
- Expected Model Change
  - モデルが一番大きく変化するデータを選択
- Expected Error Reduction
  - 学習後の”不確かさ”の期待値が最小となるデータを選択

目次

サンプリングバイアス

- 能動学習で得られたデータ集合は，実際のデータ集合と異なる
  - 当たり前だが．．
- ここで，問題が発生する
  - 能動学習のデータ集合の最適解と，実際のデータ集合の最適解がずれてしまう
  - 損失最小化問題の最適化等を行う場合は特に問題となる

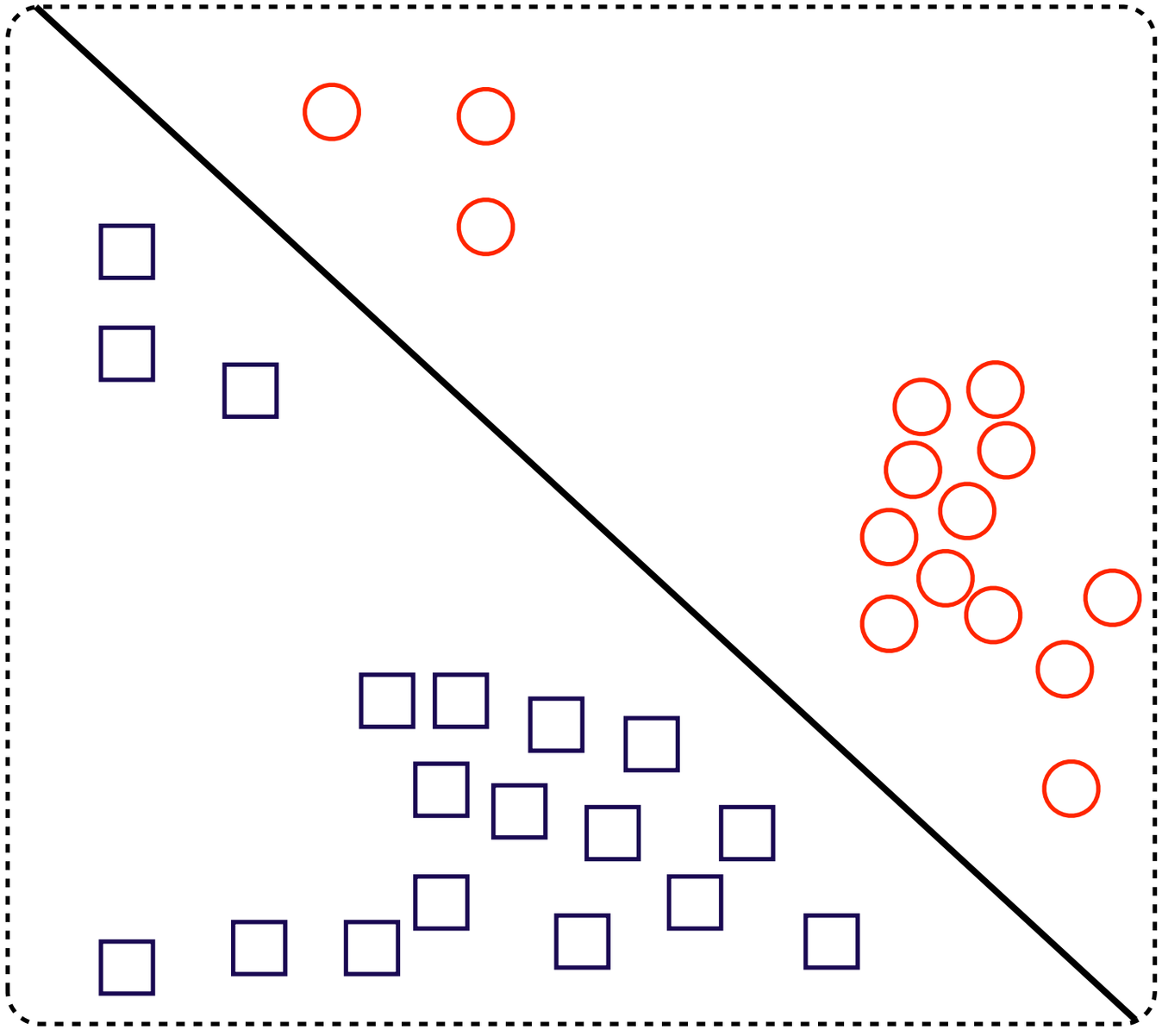
サンプリングバイアスの例



□：正例

○：負例

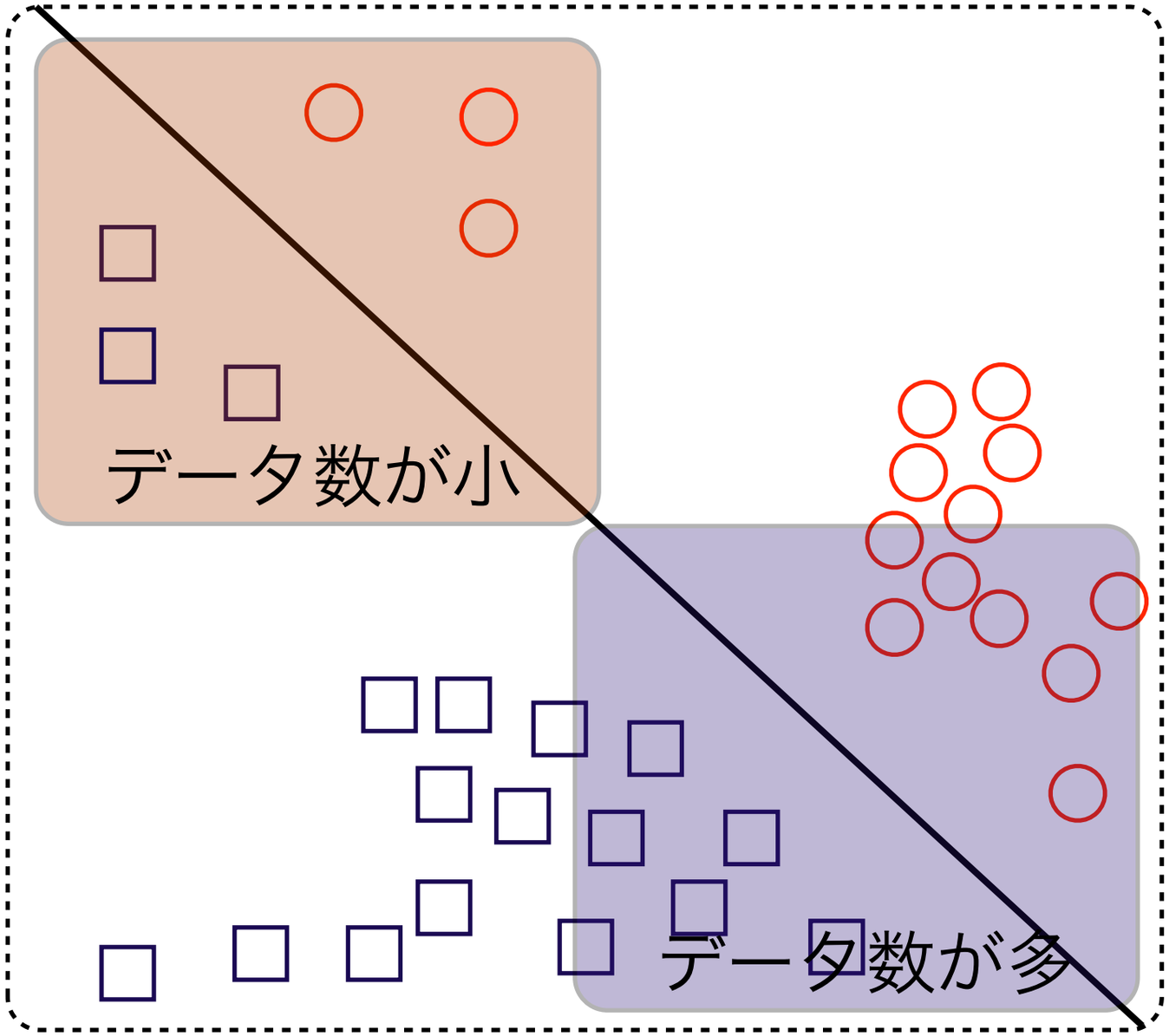
サンプリングバイアスの例



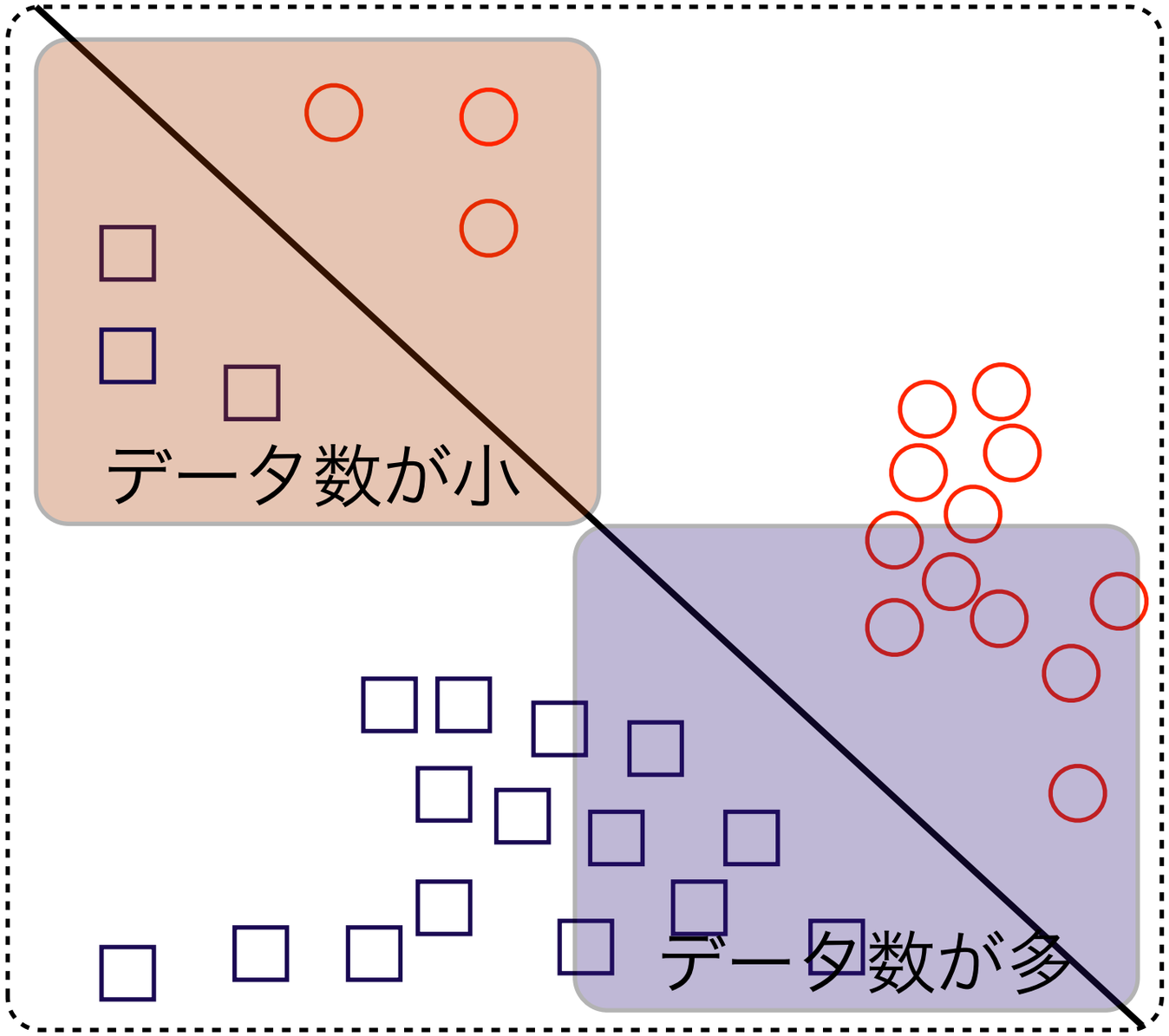
□：正例

○：負例

サンプリングバイアスの例



サンプリングバイアスの例



□：正例

○：負例

解決策

重点サンプリング

etc..

データの再利用性

- Hal Daume III の blog
  - Active Learning : far from solved

能動学習で得られるデータは，学習に用いているアルゴリズムに強く依存

- 後でアルゴリズムを変えた場合，今までに得られたデータが学習に悪影響を及ぼすことも [Baldridge+, EMNLP 2004]

データの再利用性

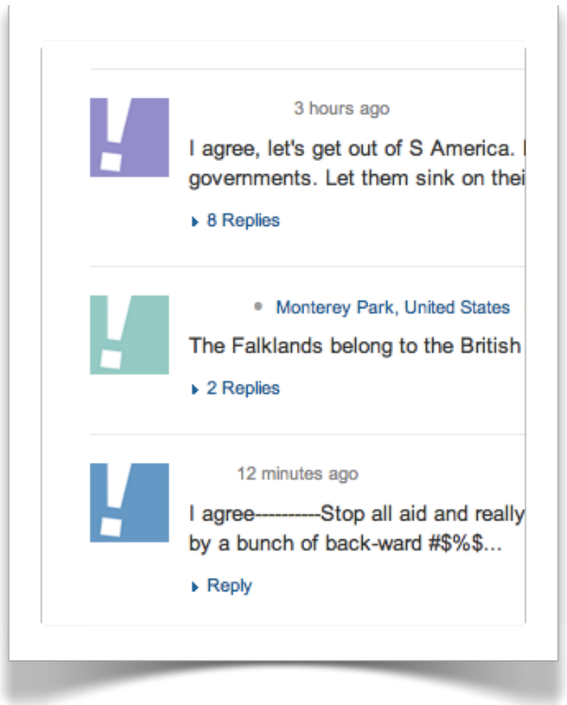
- 再利用性を増すための方法は，今のところ提案されていない
- 再利用性が問題とならないケース
  - タスクが固定
  - アルゴリズムを置き換えない
- このケース以外は，データの再利用性が弱点となる可能性

目次

Unbiased Online Active Learning in Data Stream

[Chu+, KDD 2011] (Yahoo! Labs)

- ユーザー生成型コンテンツからスパムを排除するタスク
  - ニュースサイトのコメント欄等が対象
  - 毎日，大量のコンテンツが生成
  - 最小の労力で高い精度のスパムフィルターを作りたい
- ストリームベース能動学習
  - 重点サンプリング＋エントロピー方式
  - その他幾つかのモデル拡張を提案



実験

- ニュースサイトのコメント欄30日分
  - 初日分は全て学習
  - その他は能動学習
- 提案手法は少量のデータで高い精度
  - コンセプトドリフトに自動で対応
  - 量を1/10にしても精度はあまり変化しない

データ数	26万
特徴次元数	27万
1データ当たり非零要素数	Ave. 90
スパム率	1~5%

Dualist

- http://code.google.com/p/dualist/
  - Java実装 (中にmallet)
- テキスト処理用ツール
  - 文書分類
  - 情報抽出
  - Twitterの評判分析
- 能動学習＋半教師あり学習
  - 文書と単語の二方面からラベル付け可能

Closing the Loop: Fast Interactive Semi-Supervised Annotation With Queries on Features and Instances

[Settles, EMNLP 2011]

- プールベース能動学習
  - 文書選択基準：エントロピー方式
  - 単語選択基準：Information Gain
- 多項ナイーブベイズ + EM
  - E step : 各単語のカテゴリ確率を計算
  - M step : 各文書のカテゴリ確率を計算
  - ラベル付けされた単語は，事前分布確率が上昇

KyTea

- 単語分割／読み・品詞推定のための解析器
  - 点推定と部分的アノテーション
- 能動学習（単語分割） [Neubig+, NLP2010]
  - SVMの分離平面から近いデータを選択
  - ANPI_NLPの際にも利用
  - http://www.phontron.com/kytea/active-ja.html
  - 部分アノテーションと相性が良い (自信のあるところだけアノテーションすればよいため)

参考資料

- A tutorial on Active Learning [Dasgupta, ICML 2009]
  - http://hunch.net/~active_learning/active_learning_icml09.pdf
- Active Learning Literature Survey [Settles, Techreport 2010]
  - http://active-learning.net/
- A Two-Stage Method for Active Learning of Statistical Grammars [Becher+, IJCAI 2005]
- Large Language Models in Machine Translation [Brants+, EMNLP 2007]
- Semi-Supervised Learning with Graphs [Zhu, Ph.D Thesis 2005]
- Active Learning with Support Vector Machines in the Drug Discovery Process [Warmuth+, Jour. of Chemical Information Science 2003]
- Active Learning and The Total Cost of Annotation [Baldridge+, EMNLP 2004]
- Importance Weighted Active Learning [Beygelzimer+, ICML 2009]
- Agnostic Active Learning Without Constraints [Beygelzimer+, NIPS 2010]
- 点推定と能動学習を用いた自動単語分割器の分野適応 [Neubig+, NLP 2010]

問題設定

目的：<u>経験損失の最小化</u>

$$\hat{R}_n = \frac{1}{n}\sum_{i=1}^{n} G(\mathbf{x}_i, y_i; \theta)$$

$$\{(\mathbf{x}_i, y_i)\} \sim p(\mathbf{x}, y)$$

	記法
入力	$\mathbf{x}$
出力	y
パラメータ	θ
真の確率分布	$p(\mathbf{x}, y)$
損失関数	$G(\mathbf{x}_i, y_i; \theta)$

少量のデータでサンプリングするため,

真の確率分布と異なる分布になる

可能性が高い [Sampling bias]

<u>重点サンプリング</u>

提案分布を基に,
ラベル付与の有無を決定

重点サンプリング

目的：経験損失の最小化

$$\hat{R}_{n,q} = \frac{1}{\mathcal{B}} \sum_{i=1}^{n} \frac{p(\mathbf{x}_i, y_i)}{q(\mathbf{x}_i, y_i)} G(\mathbf{x}_i, y_i; \theta)$$

$$\{(\mathbf{x}_i, y_i)\} \sim q(\mathbf{x}, y)$$

	記法
入力	$\mathbf{x}$
出力	y
パラメータ	θ
真の確率分布	$p(\mathbf{x}, y)$
損失関数	$G(\mathbf{x}_i, y_i; \theta)$
提案分布	$q(\mathbf{x}, y)$

- 提案分布と真の確率分布との比でデータを重み付け

$$\frac{p(\mathbf{x}_i, y_i)}{q(\mathbf{x}_i, y_i)} = \frac{p(\mathbf{x}_i)p(y_i|\mathbf{x}_i)}{q(\mathbf{x}_i)p(y_i|\mathbf{x}_i)} = \frac{p(\mathbf{x}_i)}{q(\mathbf{x}_i)}$$

ラベル無しデータのみから計算可能

重点サンプリング

- 不偏性を保証
  - サンプリングされたデータからの統計量が元データの統計量に一致

$$E_{\mathbf{x}\sim q}[\hat{R}_{n,q}(\theta)] = E_{\mathbf{x}\sim p}[R_n(\theta)]$$

- 提案分布の設定
  - Uncertainty Samplingのエントロピー方式等