公正配慮型分類器の公正性に関する分析

神嶌 敏弘[*1]，赤穂 昭太郎[*1]，麻生 英樹[*1]，佐久間 淳[*2]

[*1] 産業技術総合研究所，[*2] 筑波大学

第18回 情報論的学習理論と機械学習 (IBISML) 研究会

筑波大学，2014.9.1-2

開始

はじめに

[Romei+ 13]

公正配慮型データマイニング

公正性，差別，中立性，独立性などの潜在的な社会的問題について配慮しつつデータマイニングを行う

↓

公正配慮型分類

- 公正配慮型データマイニングの代表的なタスク
- ある公正性制約の下で出来るだけ正確に予測する分類器を，潜在的に不公正な決定を含む訓練データから学習する

✤ 差別配慮型DMとも呼ばれているが，ここでは公正配慮型DMと呼ぶ．これは，差別の英語 discrimination が機械学習の文脈では判別の意味になることと，差別への対処以外の問題への適用も可能であるためである．

はじめに

CaldersとVerwerの2単純ベイズ法

Calders & Verwer's 2- naive Bayes (CV2NB)

非常に単純ではあるが，有効な手法

他の公正配慮型分類器は精度で上回ることはあっても公正性は劣る

↓

CV2NBの性能が優れている理由を明らかにする

- モデルバイアス
- 確定的な決定則

↓

- CV2NB法と同等の性能で，その動作原理が明確な手法の提案
- 生成モデル型以外のどの分類器にも適用できるような拡張

目次

目次

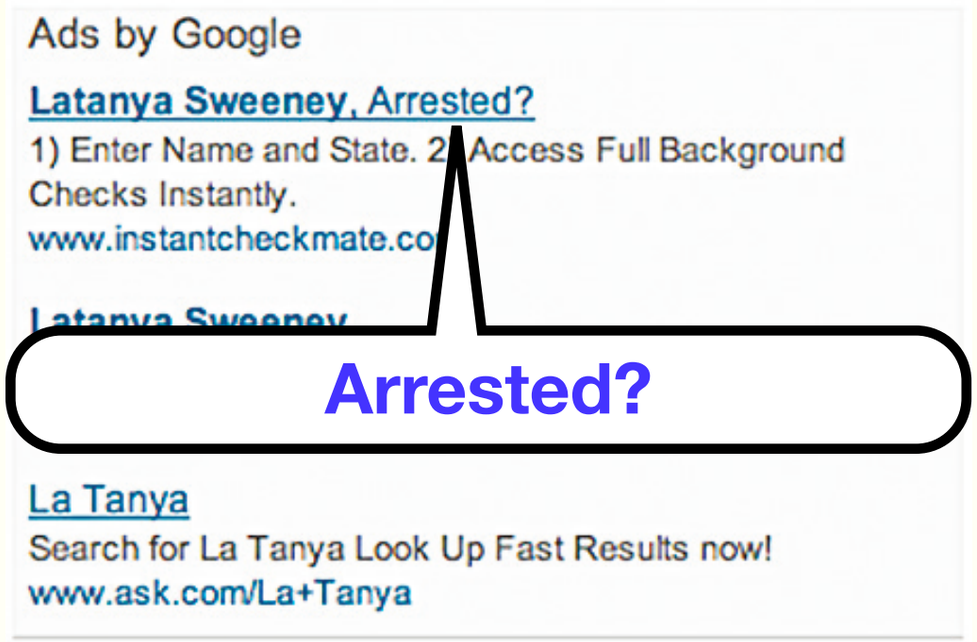
差別的決定の防止

[Sweeney 13]

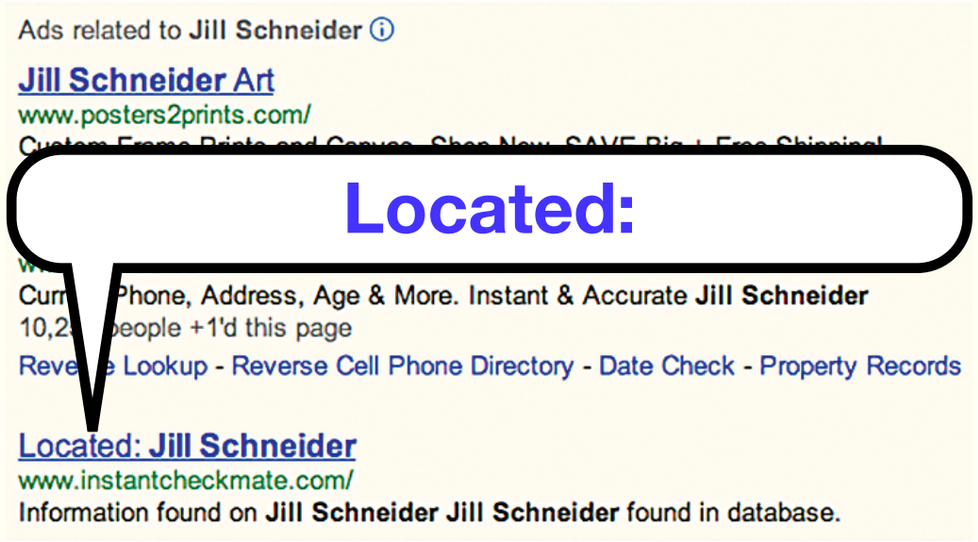
キーワードマッチ広告配信での懸念

逮捕歴を示唆するような広告文が，ヨーロッパ系で多い名前より，アフリカ系で多い名前でより頻繁に表示された

アフリカ系の名前



ヨーロッパ系の名前



対象者の人種の情報は用いておらず，クリック率向上による副次的な影響によるものであった

このような不公正な決定は公正配慮型DM技術で回避できる

中立的な情報の提供

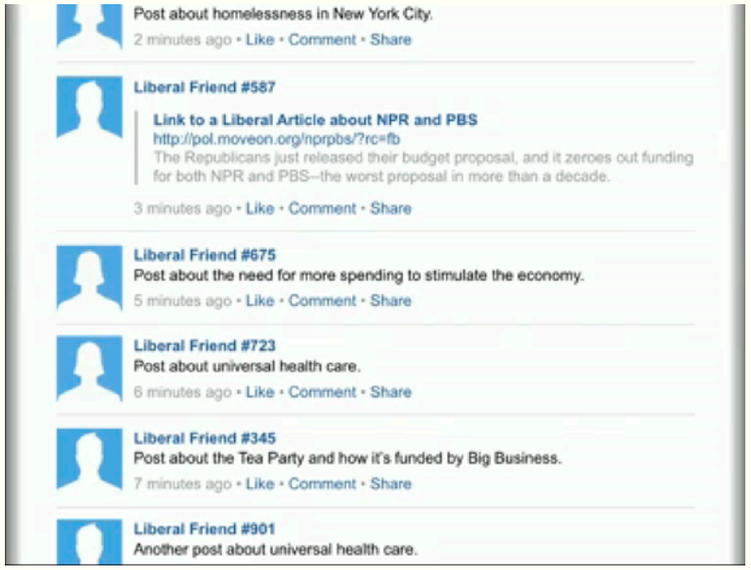
[TED Talk by Eli Pariser, http://www.filterbubble.com/]

フィルターバブル問題

Pariserは，個人化技術により，人々がふれる情報の話題に偏りが生じ，また狭まるとの懸念を示した．

Facebookの友人推薦の例

Pariserの嗜好に合わせて，友人推薦リストから保守派の人が，
知らされない間に除外されていた



FADM技術は中立的な情報を提供するのに役立つ

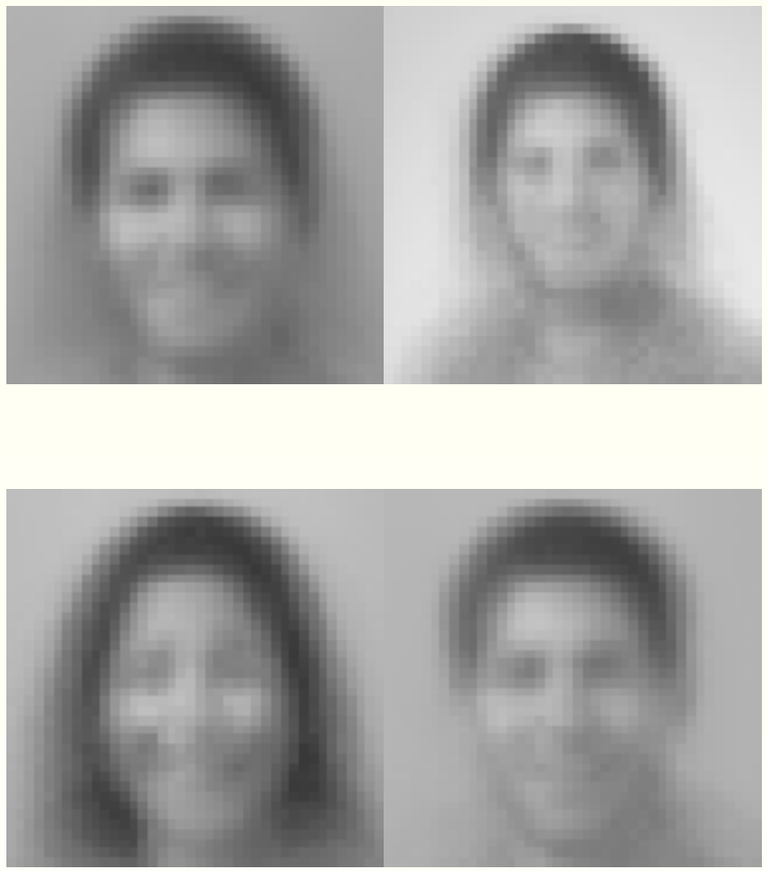
関心のない情報の除外

[Gondek+ 04]

非冗長クラスタリング (non-redundant clustering)

無関心な分割とはできるだけ独立な分割を抽出するクラスタリング

情報ボトルネック法を拡張した，条件付き情報ボトルネック法



顔画像集合のクラスタリング

- 単純にクラスタリングすると，顔だけと，肩も含めた画像に分割された
- 分析者は，こうした分割には意味的に興味深くないと考えた
- この分割とは独立となるようにクラスタリングすると男女のクラスタが得られた

FADM技術により，関心のない不要な情報を除外できる

目次

基本的な表記

Y **目的変数**
objective variable

- 重大な決定の結果
  例：ローンの可否，採用，入試

S **センシティブ特徴**
sensitive feature

- 社会的に配慮が必要な情報
  例：性別・人種

$\mathbf{X}$ **非センシティブ特徴ベクトル**
non-sensitive feature vector

- 要配慮特徴以外の特徴
- 直接の配慮は不要だが，センシティブ特徴と相関がある場合も

データマイニングにおける公正性

データマイニングにおける公正性

センシティブな情報が決定に影響しない

↓

センシティブ特徴と相関がある非センシティブ特徴は
センシティブ情報を含んでしまっている

↓

red-lining 効果：たとえセンシティブ特徴を利用せずに計算しても，
公正な決定はできない

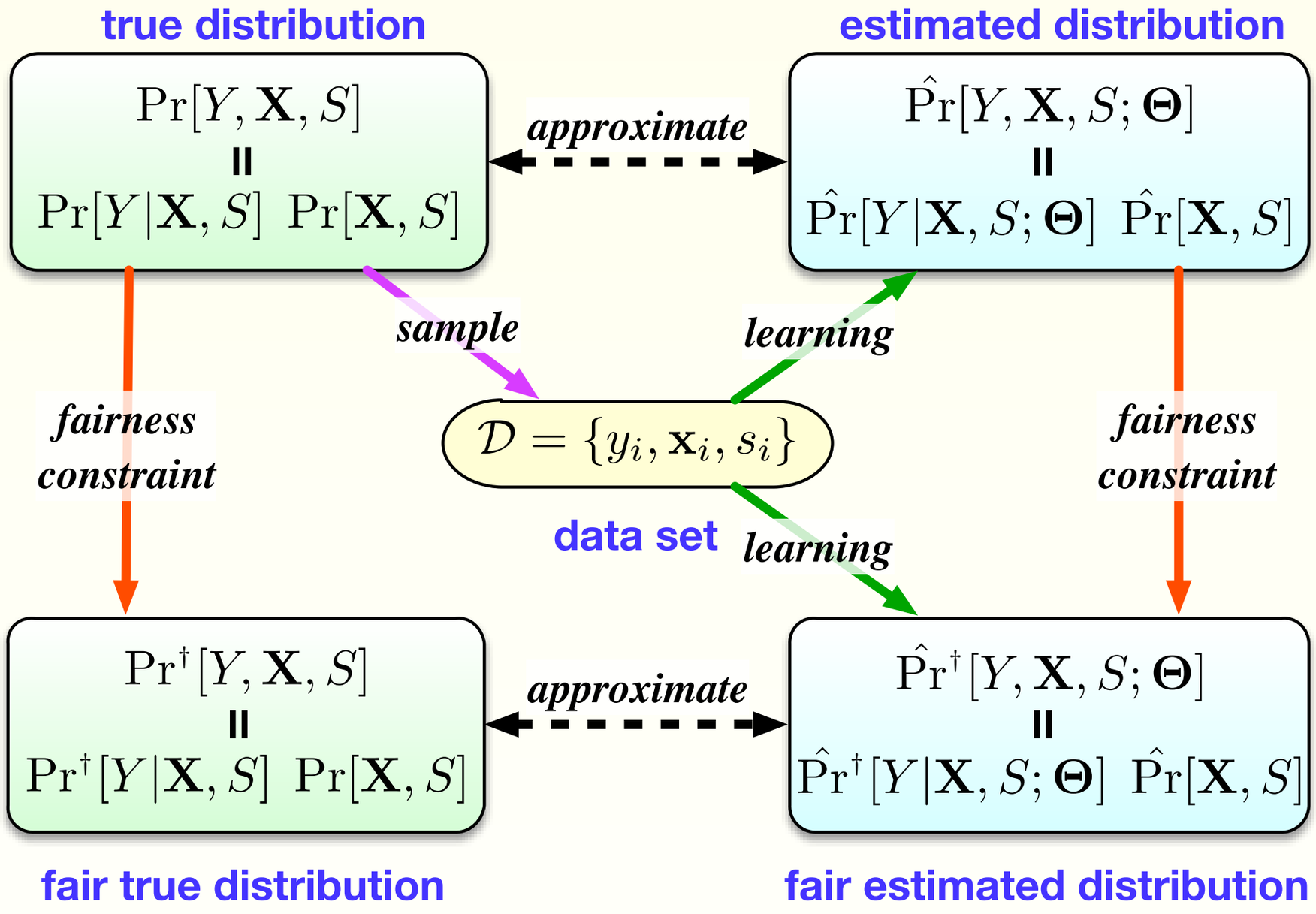
Y と S は，$\mathbf{X}$ が与えられたときの条件付き独立：$Y \perp\!\!\!\perp S \mid \mathbf{X}$

↓

センシティブ特徴と目的変数は
無条件に独立である必要 $Y \perp\!\!\!\perp S$

分布の表記

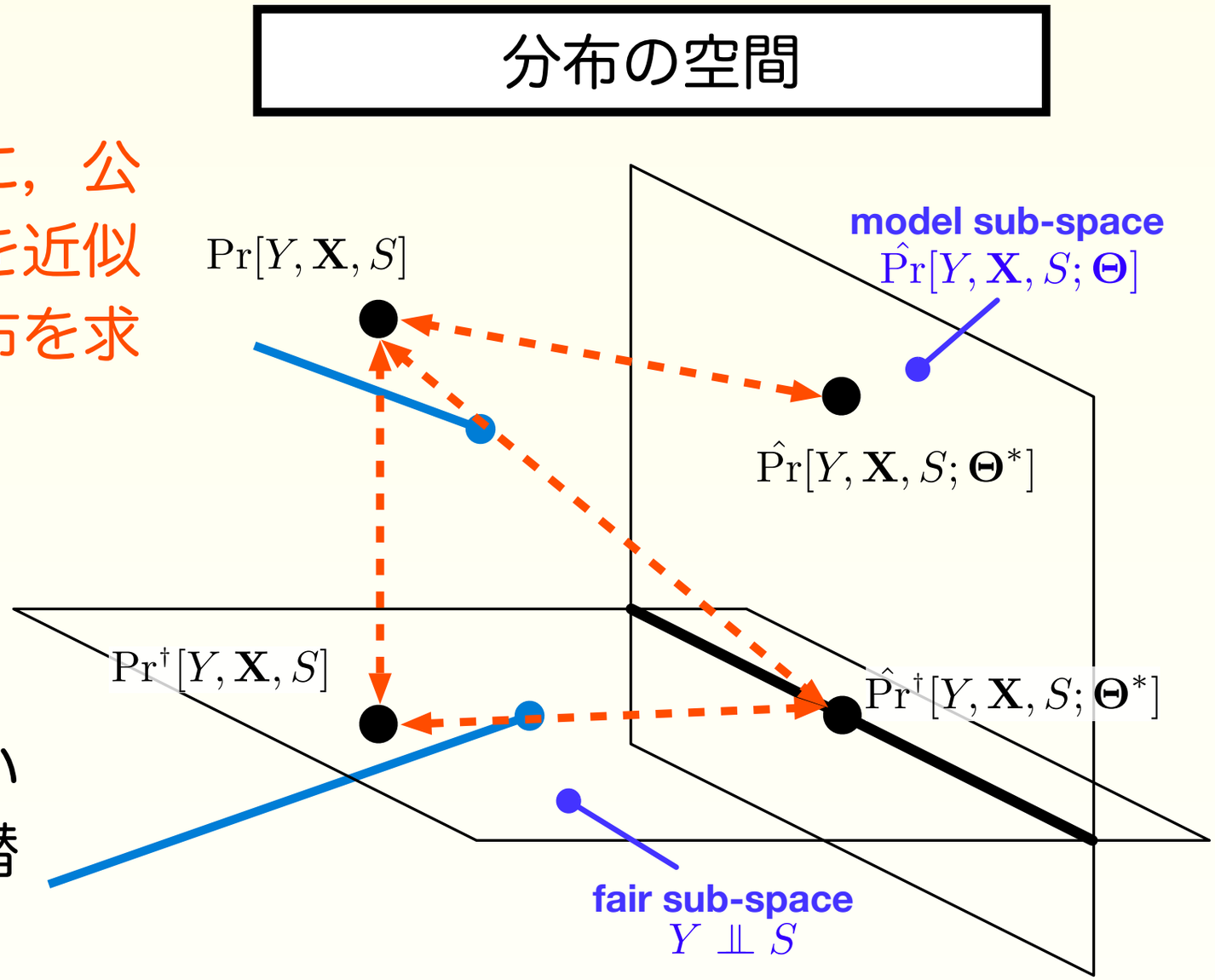
通常の真の分布と推定分布に加えて,
公正性制約を満たす公正分布（記号†で表す）も考える



公正配慮型分類

公正配慮型分類
公正な真の分布の代わりに，公正性制約の下で真の分布を近似するような公正な推定分布を求める

公正な真の分布を求めたいのだが，実世界の事例は潜在的に不公平であるため，この分布からの標本は得られない．

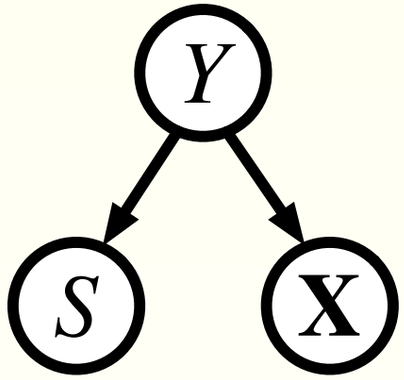


Calders と Verwer の2単純ベイズ

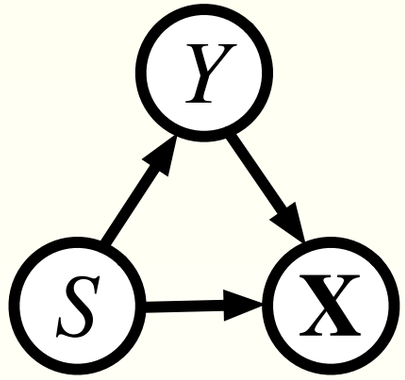
[Calders+ 10]

非センシティブな特徴 X が目的変数 Y だけでなく
センシティブ特徴 S にも依存することで不公正な決定をモデル化

単純ベイズ → 2単純ベイズ (CV2NB)



- S と $\mathbf{X}$ は，Y が与えられたときに条件付き独立



- $\mathbf{X}$ 中の各特徴は，S と Y が与えられたときにそれぞれ条件付き独立

✤ S のそれぞれの値に応じて二つの単純ベイズ分類器を学習するのと等価なため2単純ベイズ法と呼ぶ

Calders と Verwer の2単純ベイズ

[Calders+ 10]

CV2NBの予測モデル：特徴の分布は Y と S の両方に依存

$$\Pr[Y, X, S] = \Pr[Y, S] \prod \Pr[X_i | Y, S]$$

公正な決定をさせるために　$\hat{\Pr}[Y, S]$　を修正する

推定モデル $\hat{\Pr}[Y, S]$ **fair** 公正推定モデル $\hat{\Pr}^{\dagger}[Y, S]$

経験分布と予測分布が一致するように

Pr[Y, S] の修正アルゴリズム

if データの分類結果の正例 ＜ 元の訓練データの正例 **then**

　　Pr[Y=有利, S＝保護]を増加，Pr[Y=不利, S＝保護]を減少

else

　　Pr[Y=不利, S＝非保護]を増加，Pr[Y=有利, S＝非保護]を減少

更新した Pr[Y, S] を用いてデータを再分類

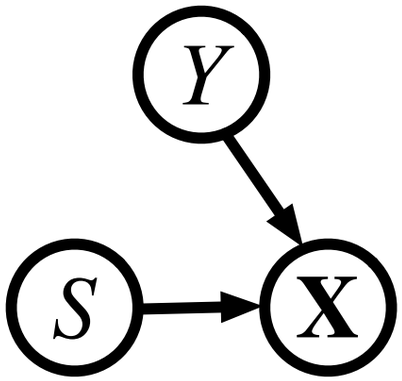
差別スコアが小さくなるように

目次

仮説公正分解

仮説公正分解

分類器を公正にするためのモデル化手法



- 分類モデルで，センシティブ特徴と目的変数とを無関係にする
- **この手法により，センシティブ特徴と目的変数を統計的に独立にする**

仮説公正分解単純ベイズ
Hypothetical Fair-Factorized Naive Bayes (HFFNB)

仮説公正分解を単純ベイズに適用

$$\hat{\Pr}^{\dagger}[Y, \mathbf{X}, S] = \hat{\Pr}^{\dagger}[Y]\hat{\Pr}^{\dagger}[S]\prod_k \hat{\Pr}^{\dagger}[X^{(k)}|Y, S]$$

単純に事例数を数えればパラメータを最尤推定はMAP推定できる

ROC決定則との関連

[Kamiran+ 12]

公正化していない場合に新規の対象 $(\mathbf{x}, s)$ をクラス1に分類する条件

$$\hat{\Pr}[Y=1|\mathbf{x}, s] \geq 1/2 \equiv p$$

↓

Kamiran らの ROC 決定則

センシティブ特徴の値に応じて決定しきい値 p を変更する

↓

HFFNB法は決定しきい値 p を次式に変更しているのと等価

$$p' = \frac{\hat{\Pr}r[Y|S](1-\hat{\Pr}[Y])}{\hat{\Pr}[Y]+\hat{\Pr}[Y|S]-2\hat{\Pr}[Y]\hat{\Pr}[Y|S]}$$

(コスト考慮型学習の Elkanの定理より)

HFFNB法は，ROC決定則を適用した場合の特殊な場合とみなせる

CV2NB 対 HFFNB

CV2NB法とHFFNB法の予測精度と公正性を比較

予測精度（正解率）
大きな値ほど
より正確に予測できている

不公正度（正規化相互情報量）
大きな値ほど
決定がより不公正になっている

	予測精度	不公正度
HFFNB	0.828	1.52×10^{-2}
CV2NB	0.828	6.89×10^{-6}

HFFNB法は，CV2NB法と同等に正確に予測できているが，
はるかに不公正な決定しかできていない

なぜ?

目次

モデルバイアス

仮説公正分解したモデルでは,
データは次の推定分布から生成されていると仮定

$$\hat{\Pr}[Y]\hat{\Pr}[S]\hat{\Pr}[\mathbf{X}|Y,S]$$

↕

真の分布から生成された新規の対象について
推定分布を用いてそのラベルを推定している

推定分布 $\hat{\Pr}[Y|\mathbf{X},S]\Pr[\mathbf{X},S]$ 真の分布

↓

モデルバイアスが大きいと二つの分布は乖離する

確定的決定則

仮説公正分解したモデルでは,
次の分布に従ってラベルが**確率的に決定**されると仮定

$$\hat{\Pr}[Y|\mathbf{X}, S]$$

一方で，実際の予測ラベルは次の規則で**確定的に決定**

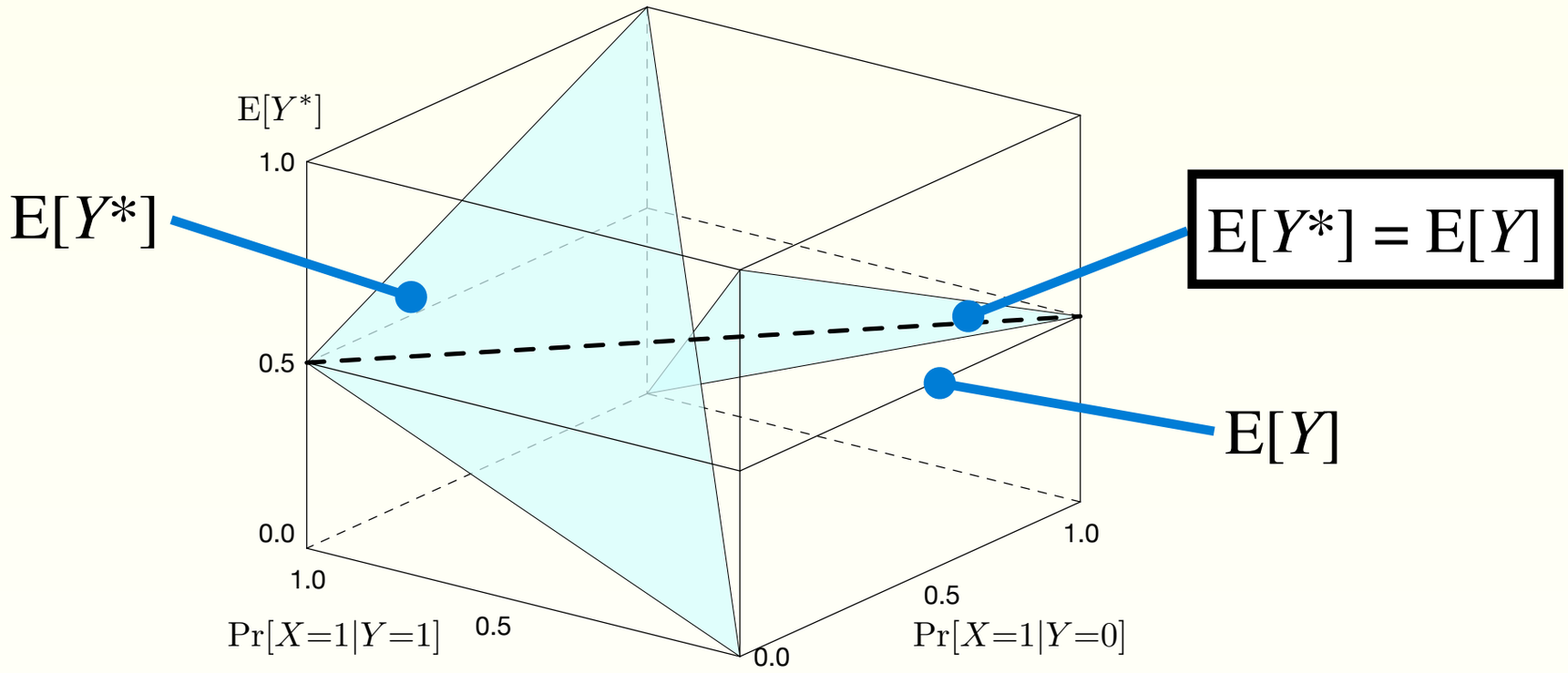
$$y^* = \arg \max_{y \in \mathrm{Dom}(Y)} \hat{\Pr}[Y|\mathbf{X}, S]$$

二つの異なる過程を経て得られたラベルは一般には一致しない

確定的決定則

簡単な分類モデル：二値クラスで二値特徴1個

- クラス分布は一様： $\hat{\Pr}[Y=1] = 0.5$
- Y^* は確定的に決定： $Y^* = \arg\max \Pr[Y|X]$
- 変更したパラメータ：$\Pr[X=1 \mid Y=1]$ と $\Pr[X=1 \mid Y=0]$



E[Y^*] と E[Y] は一般には一致しない

目次

実公正分解

HFFNB法が失敗するのは,
モデルバイアスと確定的な決定則の影響を無視しているから

↓

実公正分解 (Actual Fair-factorization)

目的変数とセンシティブ特徴を，推定した仮説上の分布ではなく，
実際に得られた分布状で無関係にする

- 仮説公正分解と同様に，クラスラベルとセンシティブ特徴とを統計的に独立にする：
- 推定分布 $\hat{\Pr}[Y, \mathbf{X}, S]$ ではなく，実際の分布 $\hat{\Pr}[Y|\mathbf{X}, S]\Pr[\mathbf{X}, S]$
- 決定的に選ばれたクラスラベルを使用する

実公正分解単純ベイズ (AFFNB)

実公正分解単純ベイズ (Actual Fair-factorization naive Bayes; AFFNB)

実公正分解を単純ベイズモデルに適用

モデルバイアス

真の分布との積　$\hat{\Pr}[Y|\mathbf{X}, S]\Pr[\mathbf{X}, S]$ を
標本平均 $(1/|\mathcal{D}|)\sum_{(\mathbf{x},s)\in\mathcal{D}}\hat{\Pr}[Y|\mathbf{X}{=}\mathbf{x}, S{=}s]$ で近似する

確定的決定則

クラスラベルの分布を使う代わりに，確定的に決定されたクラスラベルを数え上げる

Y^* と S を独立にする

Y^* と S のそれぞれの周辺分布が，対応する標本分布と等しくなるようにする

CV2NB 対 AFFNB

CV2NB法とAFFNB法の，予測精度と公正性を比較

	予測精度	不公正度
HFFNB	0.828	7.17×10^{-2}
AFFNB	0.828	5.43×10^{-6}
CV2NB	0.828	6.89×10^{-6}

CV2NB法とAFFNB法は予測精度だけでなく公正性も同等

CV2NB法の公正性が優れていたのは，センシティブ特徴とラベルの推定分布上での独立性ではなく，モデルバイアスと確定的決定則の影響を受けた実際の分布上での独立性を考えていたためであった

目次

生成モデル以外への拡張

実公正分解は，ラベルの生成確率を変更するので生成モデル専用

↓

分類のモデルは，生成モデル，識別モデル，識別関数によるもの

↓

識別モデルや識別関数による方法への拡張

↓

予測ラベルとセンシティブ特徴値に応じて確定的にラベルを決定

$$\sum_Y \hat{\Pr}^{\dagger}[Y^*=1|Y,S=s]\,\hat{\Pr}[Y|S=s]\,\hat{\Pr}[S=s]\,\hat{\Pr}[\mathbf{X}|Y,S=s]$$

↓

$$f_s^{\dagger}(\mathbf{x}) = f_s(\mathbf{x}) + b_s, \text{ for } s \in \{0,1\}$$

識別用の決定しきい値を，センシティブ特徴値に応じて変更

ロジスティック回帰とSVMでの結果

ロジスティック回帰と線形SVMに拡張実公正分解を適用した

	予測精度	不公正度
HFFNB	0.828	7.17×10^{-2}
AFFNB	0.828	5.43×10^{-6}
AFFLR	0.833	2.80×10^{-6}
AFFSVM	0.833	2.80×10^{-6}
CV2NB	0.828	6.89×10^{-6}

LRでもSVMでも，CV2NB法と同等の公正性を達成

公正性は同等になるので，データに対して予測精度の良い方法を選択すれば，よいトレードオフを達成できる

目次

プライバシ保護データマイニング

データマイニングにおける公正性

目的変数 Y とセンシティブ特徴 S の間の統計的独立性

↓

情報理論の観点からは Y と S の間の相互情報量が 0 と同値

↓

プライバシ保護の観点からは，目的変数の値が知られたときの，
センシティブ特徴の保護に該当し，t 近接性の概念に近い

プライバシ保護データマイニングとの差異

- 採用の可否など重要な決定にあたっては，ランダムに決定することが不適切な場合がある
- 個人を特定できることは公正配慮型DMでは一般には問題ではない

説明可能変数と傾向スコア

[Žliobaitė+ 11,Calders+ 13]

説明可能特徴：法的・社会通念上，決定に影響しても問題ない要因
他の条件が同じでセンシティブ特徴が違うとき，決定が異なると差別的

↓

説明可能特徴 $\mathbf{X}^{(E)}$ で条件付けた上で Y と S の独立性を考える

$$Y \perp\!\!\!\perp S \mid \mathbf{X}^{(E)}$$

説明可能特徴は Y と S の双方に影響を与える交絡因子として扱う

↓

傾向スコア：$\mathbf{X}^{(E)}$ から S=1 になる確率を予測する関数
傾向スコアで層別に分けて，各層内で公正性を保証すると，説明可能特徴の交絡因子としての影響を除去する

その他の関連分野

- **コスト考慮型学習**：公正性を損なう分類，有利な決定を受けるべき保護グループ中の個人が，不利な決定を受けると大きな誤分類コスト
- **Legitimacy / Leakage**：実世界で運用できるようなモデリング
- **独立成分分析**：特徴間の独立性を最大化するような変換
- **delegateデータ**：検定で比較するために，特定の情報を除外したデータを作成する
- **ダミークエリ**：利用者の個人情報保護のため，ダミーの検索質問や商品評価を入力する
- **Visual Anonymization**：個人を特定できないようにするために，顔などの視覚情報を削除する

まとめ

この研究の寄与

- 公正性を達成できであろう簡潔な仮説公正分解単純ベイズ (HFFNB) 法とCalders&Verwerの2単純ベイズ (CV2NB) 法とを比較し，CV2NB法の優位性を確認
- この優位性は，モデルバイアスと確定的決定則の影響を考慮しているためであるとの仮説を，実公正分解単純ベイズ (AFFNB) 法を作成することで示した
- 実公正分解法を，識別モデルや識別関数による分類器にも適用できるように拡張した

今後の予定

- 公正性の改良は，予測ラベル Y とセンシティブ特徴 S のみに基づいているが，$\mathbf{X}$ も考慮する手法を開発し，より良い公正性と予測精度のトレードオフを達成する

お知らせ

実験コードを公開しています

まだ前のバージョンですが，そのうち更新します

http://www.kamishima.net/fadm

謝辞

- 研究の詳細な情報を提供してくれた Sicco Verwer 氏，およびベンチマークデータを提供している Indrė Žliobaitė 氏に感謝する．
- 本研究はJSPS科研費 16700157，21500154，24500194，25540094 の助成を受けたものである