P R E S E N T A T I O N

Web システムにおける自動分類器を利用した機械学習と顧客購買行動の予測

Agenda

・自己紹介
後回しにします

・概説 Outline
機械学習とは、自動分類器とは、決定木とは

・解法 Algorhythm
訓練データからどのように決定木を生成していくのか

・訓練 Machine Learning
教師データを基にした機械学習

・推測 Estimation of the future
訓練された決定木による未来の予測

1. Outline

Outline

・概要
自動分類器による機械学習を利用したデータ分析方法を説明し、データの特性からユーザーの行動パターンを分析、将来動向を予測する。

・機械学習とは
人工知能（AI）の一分野であり、与えられたデータの特質を推測できる情報を生み出し、データの規則性・知識表現・判断基準を割り出し、それらを利用して将来現れるデータの予測をすることと定義する。

・自動分類器の種類
決定木、ベイジアン分類器、ニューラルネットワーク等…

・想定するシナリオ

項目	属性	説明
ユーザー名	文字列	ユーザー固有の名称で、サイトへのログイン時に付与される。
オンラインデモを見たか	Yes / No	オンラインデモにアクセスしたかどうか。
パンフレットを見たか	Yes / No	パンフレット（PDF）をダウンロードしたかどうか。
閲覧ページ数	数値	ユーザーが閲覧したページの総数。
商品を購入したか	Yes / No	結果的に商品を購入したかどうか。（※＝帰結）

ある Web サイトでのユーザーの行動と、それに対する「帰結」として商品購入の有無を想定する。不特定多数のユーザーがサイトにアクセスし、オンラインデモやパンフレットを閲覧、最終的に商品を購入する。ここで特定の行動を示すユーザーが商品を購入することを示す要素を分析すれば、広告戦略、 Web サイトの改善、優良顧客への特待販売戦略等に活用が可能である。

・購買心理過程の 8 段階

注目→興味→連想→欲望→比較検討→信頼→行動→満足

行動＝これを下さいと購買を決定する過程

2. Algorhythm

・決定木
自動分類器の一種。
学習結果をツリー状に表示できるので視覚的にわかりやすい。

・分割の評価尺度
決定木を生成するにあたり分割の基準を算出する。
帰結として取り得る値
→ここでは商品購入の有無のみ。
どの条件（たとえばオンラインデモを見たか否か）で分割するとユーザー行動を予測しやすくするかを判断する。

・不純度
不純度とは 2 集団の混合の度合い。
ジニ不純度
エントロピー不純度
他にもカイ 2 乗統計量などの評価尺度がある。

・ジニ不純度

集合中のアイテム 1 つに帰結の 1 つをランダムに当てはめる場合の期待される誤差率
ジニ係数＝もともと社会における所得分配の不平等を測定する指標としてローレンツ曲線をもとに考えられた。

・ローレンツ曲線

確率密度関数 f(x) または累積分布関数 F(x) を用いて以下の式で示される。

$$L(F) = \frac{\int_{-\infty}^{x(F)} x f(x)\,dx}{\int_{-\infty}^{\infty} x f(x)\,dx} = \frac{\int_0^F x(F')\,dF'}{\int_0^1 x(F')\,dF'}$$

・ジニ係数

0 から 1 の範囲を取り、以下の式で示される。

$$I_G(f) = \sum_{i=1}^{m} f_i(1-f_i) = \sum_{i=1}^{m}(f_i - f_i^2) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} f_i^2 = 1 - \sum_{i=1}^{m} f_i^2$$

たとえば集団の要素がすべて同じカテゴリーにあれば予測はすべて正しくなるため誤差率は 0 となる。可能な帰結が 2 つあり、すべて等しく起きるのであれば、予測が正しくない確率は 50% すなわち誤差率は 0.5 となる。

・エントロピー不純度

帰結がすべて等しければ 0 となる。混合した集団になるほどエントロピーが増大する。
→エントロピーが減少することで最少の分割であると判断できる。

$$H(X) = \sum_{i=1}^{n} p(x_i)\, I(x_i) = \sum_{i=1}^{n} p(x_i) \log_b \frac{1}{p(x_i)} = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i),$$

エントロピーを求めるにはそれぞれの要素の頻度を計算する。
p(i) = 頻度(帰結) = 出現回数(帰結) / 出現回数(行)
エントロピー H(X) = すべての帰結の p(i) x log(p(i)) の合計である

可能な帰結が 2 つあり、すべて等しく起きるのであれば E(N)=0 、予測が正しくない確率 50% すなわち誤差率 0.5 のとき最も不純であり E(N)=1 となる。
たとえば六面体のダイス（サイコロ）を振ったとき、偶数という帰結に至る確度をエントロピー不純度で示すと 1 である。

・情報利得

最適な分割基準としていずれを採用しても良いが、ここではエントロピー不純度を利用する。まずは集合全体のエントロピーを計算、次に各属性の取り得る値によってグループを分割し、再帰的にエントロピーを計算する。
ここで、最も優れた分割になる属性を計算するために 2 つの確率分布の差異を測定する尺度として情報利得（※カルバック・ライブラー情報量）を利用する。

情報利得とは集合全体のエントロピーから分割後の 2 集団のエントロピーの加重平均を引いたものである。
※ A で分岐することによって得られる情報量
Gain(A) = I(p,n) – E(A)

・加重平均

観測値に重みを付けて取る平均であり、類似度スコアに基づく数値予測をおこなう際に利用する。次の式で示される。

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}.$$

今回はすべての属性について情報利得を計算し、これが最も大きくなるものを計算する。

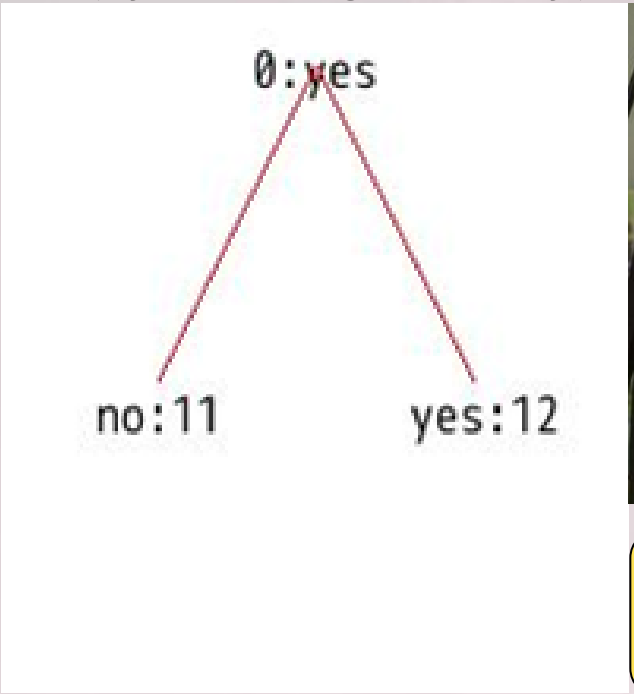
3. Machine Learning

例題（1）
不純度の低い問題

項番	ユーザー名	オンラインデモ	パンフレット	閲覧ページ数	帰結（商品購入の有無）
1	鈴木	No	No	18	No
2	田中	No	Yes	23	No
3	加藤	Yes	Yes	45	Yes
4	野村	Yes	No	33	Yes
5	佐藤	No	No	22	No
6	山田	No	Yes	19	No
7	川村	No	No	17	No
8	千反田	Yes	No	39	Yes
9	折木	Yes	No	31	Yes
10	副部	No	Yes	44	No
11	伊原	Yes	Yes	52	Yes
12	十文字	No	Yes	22	No
13	入須	Yes	Yes	88	Yes
14	本郷	Yes	Yes	34	Yes
15	中城	No	No	12	No
16	羽場	No	No	8	No
17	沢木口	Yes	Yes	38	Yes
18	海藤	Yes	Yes	17	Yes
19	杉村	No	Yes	22	No
20	遠垣内	Yes	No	30	Yes
21	陸山	Yes	Yes	39	Yes
22	河内	Yes	Yes	10	Yes
23	湯浅	No	No	27	No

例題（1）の不純度が低いことは一目瞭然である！

オンラインデモを閲覧したユーザーは必ず商品を購入するという帰結に至っている。
PIL（Python Image Library）にて学習結果から決定木を描画



枝の先頭 = オンラインデモ（項番 0）が yes かどうか
枝の末尾 = 帰結（商品購入の有無）= yes が 12 名、 no が 11 名

例題（1）の決定木生成時における推論過程の追跡

例題（2）
不純度の高い問題

項番	ユーザー名	オンラインデモ	パンフレット	閲覧ページ数	帰結（商品購入の有無）
1	鈴木	No	No	18	No
2	田中	No	Yes	23	No
3	加藤	Yes	Yes	45	Yes
4	野村	Yes	No	33	No
5	佐藤	No	No	22	No
6	山田	No	Yes	19	No
7	川村	No	No	17	No
8	千反田	Yes	No	39	Yes
9	折木	No	No	31	No
10	副部	Yes	Yes	44	No
11	伊原	Yes	Yes	52	Yes
12	十文字	No	Yes	22	No
13	入須	Yes	Yes	88	Yes
14	本郷	Yes	Yes	34	Yes
15	中城	No	No	12	No
16	羽場	No	No	8	No
17	沢木口	Yes	Yes	38	Yes
18	海藤	Yes	Yes	17	Yes
19	杉村	No	No	22	Yes
20	遠垣内	Yes	No	30	No
21	陸山	Yes	Yes	39	No
22	河内	Yes	Yes	10	No
23	湯浅	No	No	27	No

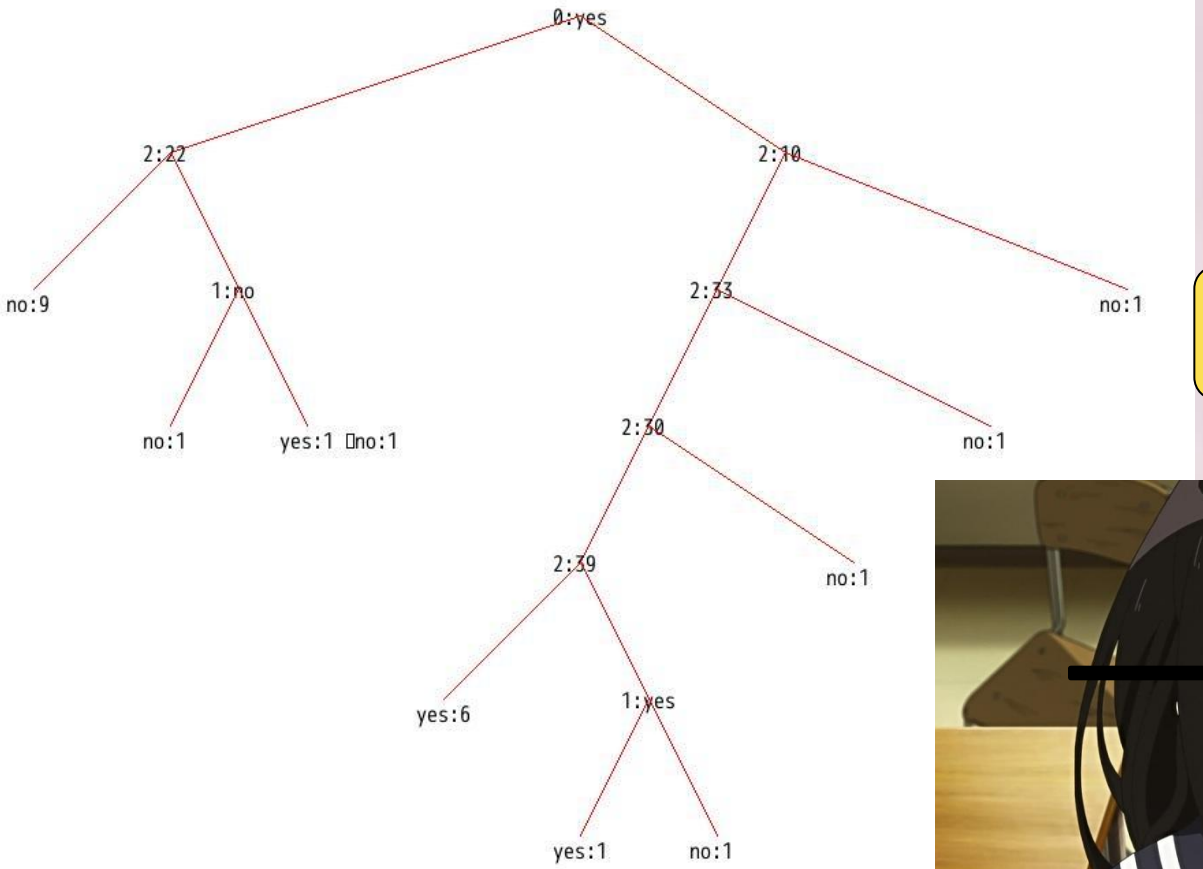
例題（2）の顧客行動と帰結の関連性は直観的な推測が難しい

オンラインデモ、パンフレット、閲覧ページ数のいずれも帰結に直接関与していない。

鈴木＝オンラインデモ:No　パンフレット:No　→ 商品を購入せず
加藤＝オンラインデモ:Yes パンフレット:Yes → 商品を購入した
副部＝オンラインデモ:Yes パンフレット:Yes → 商品を購入せず
杉村＝オンラインデモ:No　パンフレット:No　→ 商品を購入した

入須＝88ページ閲覧 → 商品を購入した
伊原＝52ページ閲覧 → 商品を購入した
羽場＝ 8ページ閲覧 → 商品を購入せず
中城＝12ページ閲覧 → 商品を購入せず
野村＝33ページ閲覧 → 商品を購入せず
副部＝44ページ閲覧 → 商品を購入せず

例題（2）による決定木
枝 = オンラインデモ（項番 0)、パンフレット（項番 1）、閲覧ページ（項番 2）
枝の末尾 = 帰結（商品購入の有無）：人数



木になります！

例題（2）の決定木生成時における推論過程の追跡

・過剰適合と正則化

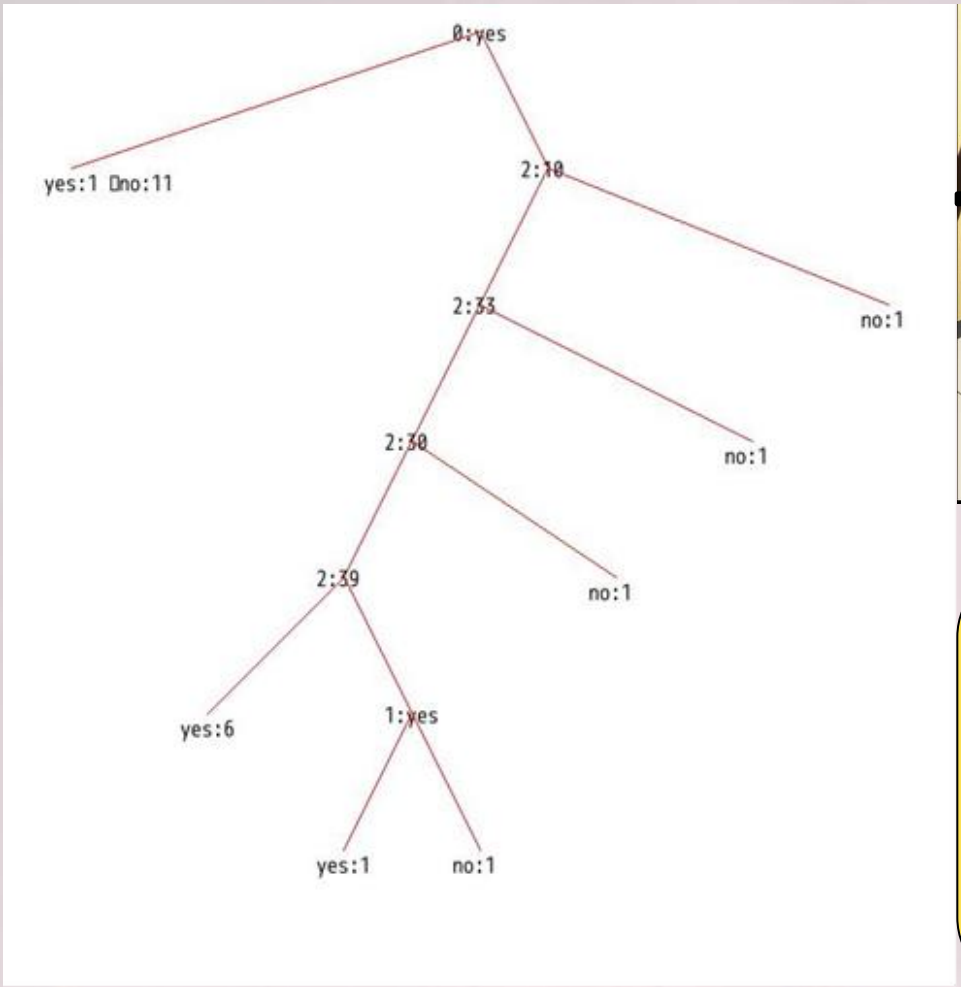
統計学や機械学習において、統計モデルの適合の媒介変数が多い場合などに、訓練データに対して学習され過ぎてしまう現象。機械学習においては過学習ともいう。
決定木でいうと評価対象の多様性に応じてツリーの分割をし続けてしまい、未知のデータに対する予測の精度をかえって下げしまう場合がある。

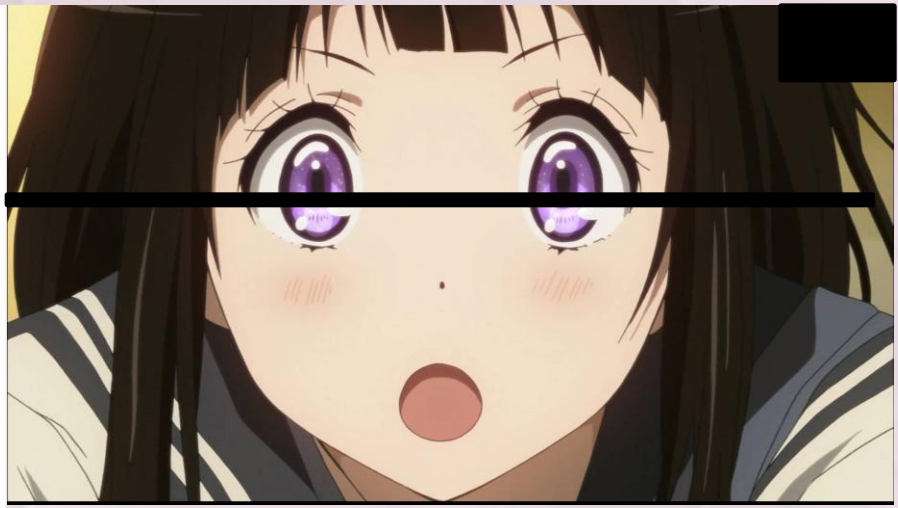
モデルの複雑さとデータ適合度のバランスを取るためには、エントロピーが全く減少しないところまで分割を続けるのではなく、情報量基準等に基づき過剰なツリーの刈り込み（※オッカムの剃刀）をする。

媒介変数を減らすのではなく、誤差関数に正則化項を追加して過学習を防ぐ。

ここではブランチ間を合成した際のエントロピー上昇の最少利得を 0.8 と設定する。
これにより閾値以下のブランチは可能な帰結をすべて含んだ単一ノードとなる。

例題（2）においてノード間合成におけるエントロピー上昇閾値を 0.8 とした決定木
枝 = オンラインデモ（項番 0)、パンフレット（項番 1）、閲覧ページ（項番 2）
枝の末尾 = 帰結（商品購入の有無）：人数





ブランチを合成した際のエントロピー上昇時最少利得を 0.8 としたときにどうなるのか、木になります！

4. Estimation of the future

・新規ユーザーの商品購入可能性

まずは適当なサンプルで実験

項番	ユーザー名	オンラインデモ	パンフレット	閲覧ページ数	予測される帰結
1	山西	Yes	No	39	Yes
2	瀬之上	No	Yes	28	No

帰結を伏せてパラメータを与えると、予測される帰結が返る。

・例題（2）をもとにした予測実験

訓練された決定木に対し、帰結を伏せて例題の予測をさせてみる。

項番	ユーザー名	オンラインデモ	パンフレット	閲覧ページ数	実際の帰結	決定木が予測した帰結
1	鈴木	No	No	18	No	No
2	田中	No	Yes	23	No	No
3	加藤	Yes	Yes	45	Yes	Yes
4	野村	Yes	No	33	No	No
5	佐藤	No	No	22	No	No
6	山田	No	Yes	19	No	No
7	川村	No	No	17	No	No
8	千反田	Yes	No	39	Yes	Yes
9	折木	No	No	31	No	No
10	副部	Yes	Yes	44	No	No
11	伊原	Yes	Yes	52	Yes	Yes
12	十文字	No	Yes	22	No	No
13	入須	Yes	Yes	88	Yes	Yes
14	本郷	Yes	Yes	34	Yes	Yes
15	中城	No	No	12	No	No
16	羽場	No	No	8	No	No
17	沢木口	Yes	Yes	38	Yes	Yes
18	海藤	Yes	Yes	17	Yes	Yes
19	杉村	No	No	22	Yes	No
20	遠垣内	Yes	No	30	No	No
21	陸山	Yes	Yes	39	No	No
22	河内	Yes	Yes	10	No	No
23	湯浅	No	No	27	No	No

・なんか 1 件まちがっているんだけど！？

23 件中 1 件の予測に失敗。
失敗したサンプルである杉村はオンラインデモもパンフレットも見ておらず閲覧ページ数も 22 ページである。同様にデモもパンフレットも見ておらず 22 ページしか閲覧していない佐藤も購入していない。

すでに杉村の行動を学習しているから予測できるのでは？
→佐藤の行動と杉村の行動は帰結を除けば全く同様で、帰結は Yes と No が 1 件ずつ
→情報利得が少ないため刈り込まれており、おおむね妥当と思われる帰結を推測

つまり杉村が商品を購入したのは意外であった

・断片的データからの行動予測

決定木では両方の枝を辿り、枝の帰結によって重み付けをして判断するという推測も可能である。
これは項目の一部が欠損している場合に有効である。たとえばログデータのうち、ブラウザが未知、リファラが不明、アクセス元が特定不能といったケースが想定される。

両方の帰結をもとに枝の重み付けを返す推測関数を追加し、欠損データから帰結を予測させてみる。

項番	ユーザー名	オンラインデモ	パンフレット	閲覧ページ数	決定木が予想した帰結
1	勝田	不明	Yes	10	No（確率 60%）
2	江波	Yes	不明	33	Yes
3	鴻巣	No	No	不明	No（確率 90%）

確率については枝の帰結から単純に重み付けをおこなった。
推測関数の実装については根拠となる理論を研究する必要がありこの辺りは課題である。

Bibliography

・参考文献

Programming Collective Intelligence: Building Smart Web 2.0 Applications
http://www.amazon.com/dp/0596529325/

集合知プログラミング
http://www.amazon.co.jp/dp/4873113644

My website

http://id774.net

@twitt

質問をどうぞ