機械学習について　　**佐藤秋彦**

機械学習とは

*「過去の**データ**より法則を見出し（**学習**し）て、**決定や分類、予測**を行うアルゴリズム」*

実社会で使われている機械学習の例：

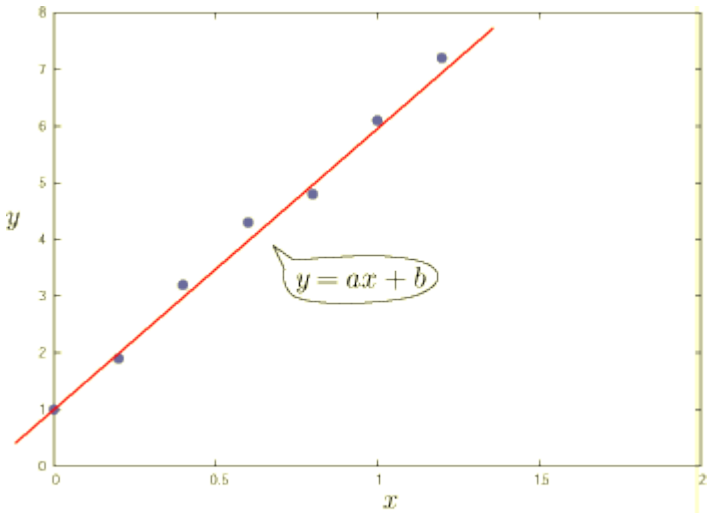
例 1.　アマゾンのおすすめ商品

多数のユーザーの**過去の購買データ**から、合わせて売れる商品の**傾向を学習**し、特定ユーザーへのおすすめ商品を**決定**する。

例 2.　空き巣犯の予知検出

人の行動を多数録画し、**動作をパラメータ化したデータ**から、空き巣犯の**動作傾向**を**学習**し、今カメラの前に来た人が空き巣犯か否かを**予測**する。

――最も原始的な先祖は線形回帰か？



（図は物理のかぎしっぽ　より） 

確かに線形回帰は**過去のデータ**（青点）より**法則**（赤線）を作成し、**予測**を行っている。
これは手計算でも求められるが、より複雑な関数でのフィッティングの場合はコンピュータの計算力を借りたくなるであろう。

線形回帰に代表される最小二乗法の他にも多数の**法則**作成方法が開発されている。具体的手法については後述するが、機械学習には大きく分けて二つの手法がある。

1.教師あり学習

性質と答えの分かった「例題」を多数読み込み、規則を見出すこと。性質は分かるが答えの分からない「本番問題」の答えを予測する。2017 年 5 月現在、佐藤はとりあえずここまで理解した。

2.教師なし学習

性質が分かっているが、特定の答えを持たないデータ群を処理する。

例：様々なパラメータが明らかになっている 10000 個の細胞を、性質の似た個体同士で 3 群に分類する。

コラム：よくニュースなどで語られる類似語

1.ディープラーニング

機械学習の一種。性質がパラメータ化されていないものの性質をパラメータ化するステップを含む。佐藤はまだ出来るようになっていない。

2.データマイニング

膨大な過去データから本質的・高価値な情報を抽出すること。データマイニングは機械学習によりなされるが、機械学習はデータマイニング以外も可能である。例えば、キノコの形質データから毒か毒でないかを判断する基準（クリティカルなのは色か？傘の形状か？など）を選定する作業はデータマイニングであるが、まだ食べたことのないキノコに対して毒か毒でないか予測をすることはデータマイニングではない機械学習である。

機械を使わず勘や経験によって過去のデータを読み解く場合、それは機械学習ではないがデータマイニングと言えるだろう。

佐藤がこれまでに身に着けた機械学習アルゴリズム集

1.最近傍法

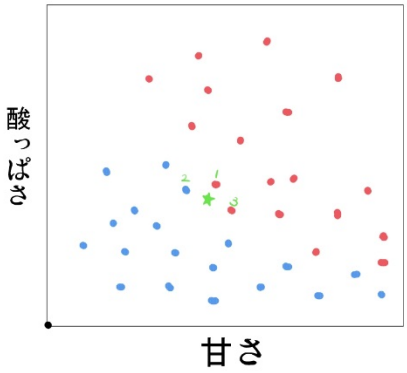
概要

複数のパラメータを持つデータの属性を決定することを目的とする。
まず、属性が分かっている複数のパラメータをもつデータ(インスタンス)を複数用意する。これが「例題」となる。
そして、属性を知りたい、属性以外のパラメータが分かっているインスタンスを「本番問題」とする。

本番問題のインスタンスと、各例題のインスタンスの(ユークリッド・マンハッタンなどの)距離を求める。

今属性を知りたい本番問題インスタンスから最も距離の近いk個の例題インスタンスの属性で多数決を行い、本番問題インスタンスの属性を予測する(kは多すぎても良くない)。

イメージ図



問題:赤点は果物、青点は野菜であることが既知のデータである。
では、緑の点は?

答え:近傍3点のうち、赤は2、青は1なので、赤の仲間。

実際用いるデータの例

細胞が癌化しているかどうかの判定を行う。

例題データ:

直径の時間平均=12.32 um, 直径の標準偏差=0.236, 周囲長の時間平均=78.85 um, 対称度=0.195, …など、30個のパラメータ
↓
答え:癌

他にも、癌であることがわかっている細胞のデータと癌でないことがわかっている細胞のデータを500データ程集める。

本番問題データ:

それでは、
直径の時間平均=10.32 um, 直径の標準偏差=0.462, 周囲長の時間平均=80.83 um, 対称度=0.225, …など、30個のパラメータ
↓
これは癌?癌でない? → 最も近い3つの例題データがそれぞれ癌・癌・癌でないなので、癌と推測!

評価

良い点：シンプルだが、データによってはそこそこ高い正答率を持つ。
悪い点：学習していない（全体的な傾向をつかもうとしていない）ので、教師（例題）データを増やすことではそれほど精度の向上が図れない。

2.単純ベイズ法

概要

学習用データより、原因→結果の条件付き確率を調べる。ベイズの定理を用いて、結果→原因の予測を行う。

ベイズの定理

$$P(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)}$$

$P(A)$は、「A が起こる確率」であり、$P(B|A)$は「B が起こったときに A が起こる条件付確率」である。今回は A を結果、B を原因とする。
すると、左辺は結果 A が起こったとき、原因 B が起こる（起こっていた）確率である。

例：高収入男性と美女のカップル（以下スーパーカップルと呼ぶ）が結婚式を挙げる確率は 96%、そうでないカップルが結婚式を挙げる確率は 30%というデータがあった。それを見た私の友人（挙式済み）は「じゃあ、私たちって世間的に見れば高収入男性と美女のカップルだったのかも！」と言った。これはどれくらいの確率なのだろうか？？
この場合、原因は「スーパーカップルである」であり、結果は「結婚式を挙げる」である。

$$(\text{挙式したカップルがスーパーカップルである確率})=\frac{(\text{あるカップルがスーパーカップルである確率})(\text{スーパーカップ}}{(\text{カップルが挙式する確率})}$$

例

メールによく出てくる 100 個の言葉$(W_1, W_2, W_3 \ldots)$があるメールの中に在るか無いかを見て、そのメールがスパムであるかどうか判定する。
例題データは、スパムかハムか判定済みで、$W_1, W_2, W_3 \ldots$がそれぞれ使われているか否か判明している（テキスト検索をする）。本番問題データは、$W_1, W_2, W_3 \ldots$がそれぞれ使われているか否かは判明しているが、スパムかハムかは判明していない。
今、$W_1, \neg W_2, W_3, \neg W_4 \ldots$というメールがあったとする。このメールがスパムである確率を計算する。

$$P(\text{スパム} \vee W_1 \cap \neg W_2 \cap W_3 \cap \neg W_4 \ldots)=\frac{P(\text{スパム})P(W_1 \cap \neg W_2 \cap W_3 \cap \neg W_4 \ldots \text{である確率} \vee \text{スパム})}{P(W_1 \cap \neg W_2 \cap W_3 \cap \neg W_4 \ldots)}$$

ただし、$W_1, W_2, W_3 \ldots$がメール中に出現するか否かは独立事象であると仮定し、$P(W_1 \cap \neg W_2 \cap W_3 \cap \neg W_4 \ldots)=P(W_1)P(\neg W_2)P(W_3)\ldots.$とする。
同様に、スパムメール中に$W_1, W_2, W_3 \ldots$が出現するか否かも独立事象であると仮定し、$P(W_1 \cap \neg W_2 \cap W_3 \cap \neg W_4 \ldots \text{である確率} \vee \text{スパム})=P(W_1 \vee \text{スパム})P(\neg W_2|\text{スパム})P(W_3|\text{スパム})P(\neg W_4|\text{スパム})\ldots$とする。
$P(W_1)$や$P(W_1|\text{スパム})$はそれぞれ、「全メールでW_1が出現する確率」と「スパムメール中でW_1が出現する確率」であり、例題データから学習することが可能だ。

評価

悪い点：$W_1 \cap W_2 \cap \neg W_3 \ldots\ldots = P(W_1)\,P(W_2)\,P(\neg W_3) \ldots.$と近似してしまっている。つまり、$W_1, W_2, W_3 \ldots$を独立な事象と見做してしまっているのだ。先の例で言えば、実際には”購入”という単語が出ることと”お金”という単語が出ることとは従属であろう。

良い点：悪い点と表裏一体であるが、条件の ON-OFF をそれぞれが独立と仮定しているので、計算処理が速い点が挙げられる。仮にそれぞれの相関まで考えるとしたら、独立を仮定した場合の約$\frac{2^{100}}{100}$倍のメモリが必要になってしまうのだ。

3.決定木モデル

概要

例題データの持つフィーチャーに関して最も効率の良い質問を次々行い、例題データを分類していく。
その後、フィーチャーのみがわかる本番問題データがどの分類先に行くかによって、本番問題データの属性を決定する。

イメージ:アキネイター
様々な人物・キャラクターの特徴をデータ化し、プレイヤーがあらかじめ想像した人物(本番問題)の属性(名前)を決定するのに最も効率の良い質問を次々に投げかける。

まずプレイヤーはダース・ベイダーを想像しておく。

女性?:いいえ → 架空の人物?:はい → 物語の主人公?:部分的にそう → (中略)
→ マントをつけている?:はい → ダース・ベイダー

今回は15問で正解(分類先に1つしか該当者がいない状態)にたどり着いた。

“最も効率の良い質問”とは?

情報学で言うエントロピーとは、熱力学的なエントロピーと同様に乱雑さを示す値である。この乱雑さを最も減らせる質問こそが最も効率の良い質問である。ただし、アキネイターとは違い、決定木モデルでは2択の質問のみを使う。
ある集団(セグメント)Sをc個のクラスに分けたとき、構成員(インスタンス)がi番目のクラスである確率をp_iとおくと、

$$\text{エントロピー}(S)=\sum_{i=1}^{c}-p_i\log_2(p_i)$$

と定義できる。例えば、赤玉が60%・白玉が40%入っている箱のエントロピーは
$-0.6\log_2(0.6)-0.4\log_2(0.4)=0.97$
一方で赤玉が100%の箱や白玉が100%の箱ならば、エントロピーは
$-1\log_2(1)=0$
である。最初の箱の赤玉と白玉を色ごとに分けた場合、分ける前と後のエントロピーの差は
$0.97-(0+0)=0.97$
となる。これを**情報獲得**と呼ぶ。「分割により乱雑さ(エントロピー)が大幅に減った」=「より分類整頓されて情報が明確になった」と解釈

できるのだ。
一方で、同じ箱の中の玉には色とは独立して、オレンジ味 80%とリンゴ味 20%がある。箱の中の玉を味に注目してオレンジ味とリンゴ味に分けるとしたら、情報獲得は
$(-0.8\log_2(0.8)-0.2\log(0.2))-(0+0)=0.72$
である。
味で分類して分割するよりも色で分類して分割する方が、**情報獲得が多い＝より効率的な分割**と判断する。

さらに、同じ箱の中の玉には色や味とは独立して、鉄製 50%とアルミ製 45%と木製 5%がある。箱の中の玉を材質に注目して鉄製とアルミ製木製の 2 パーティションに分けるとしたら、情報獲得は
$(-0.5\log_2(0.5)-0.45\,log_2(0.45)-0.05\,log_2(0.05))-¿$
となる。素材による分割は、味による分割よりは効率的で、色による分割より非効率的である。

4.線形回帰モデル

概念

間隔尺度（差に意味を持つ量的変数）を答えに持つ例題を線形回帰を複数回用いて学習し、本番問題に解答する。
フィーチャーは原則として間隔尺度が用いられるが、名義尺度はダミー変数として処理することができる。

データ例

アメリカ在住の個人の医療費を予測する。

例題データ：

年齢=19, 性別=女, bmi=27.9, 子供の数=0, タバコ=吸わない, 住所=northeast
↓
答え：16885ドル

他にも医療費と上記項目が分かっているデータを 1000 件

本番問題データ：

年齢=23, 性別=男, bmi=23.9, 子供の数=2, タバコ=吸わない, 住所=northeast
↓
答え：?ドル

評価

良い点：パラメータが少なくても高い精度を得られる。何より、具体的な量的変数を予測してくれるのがありがたい。
悪い点：間隔尺度・比例尺度でないパラメータを（原則として）使うことができない。また、あるパラメータがあったとして、それの1乗・2乗・N 乗・指数関数・対数関数などのうちどれが最も線形回帰しやすそうかを事前にある程度予測しておく必要がある。

他にも、線形回帰と決定木をハイブリッドさせたモデルなどがある。