

知的情報処理

4. ナイーブなベイズ法

櫻井彰人
慶應義塾大学理工学部

生成モデルのベイズ推定

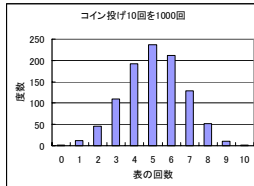
$$p(m|x) = \frac{p(x,m)}{p(x)} = \frac{\overbrace{p(x|m)}^{\text{条件付き確率}} \overbrace{p(m)}^{\text{事前確率}}}{\underbrace{p(x)}_{\text{事後確率}}}$$

を用いて、 x が観測されたとき、最もありうる m をその事後分布を用いて推定することを考える。

- $p(m)$ はクラス m の頻度で推定すればよい
- では、 $p(x|m)$ はどうしたら推定できるだろうか？
- その問題を解決した一つの方法が naïve Bayes法

属性数が問題

- 前のスライドの図を思い出してください。
- コイン投げをして、コインの表が出る確率を推定する問題と考えてください。
- 正解は 0.5 です。
- しかし、10回投げたうち、3回以下しか表が出ない場合が169回、7回以上出てしまう場合が192回もある。
- つまり、2値属性のパラメータを1個推定するにも、サンプル10個では不足だということである。
- 独立な属性が10個あれば、それらのパラメータをまあまあの精度で推定するには、 $10 \cdot 2^{10} = 1$ 万個のサンプルが必要になる。
- 一般にはなかなか難しい。



では、どうするか？

定式化: Bayes 推論使用時の課題

- サンプルの属性 $x = \langle a_1, \dots, a_n \rangle$ が与えられたとき、 x が属するクラス v を最尤推定するには？

$$v_{MAP} = \arg \max_{v_j \in V} P(v_j | a_1, a_2, \dots, a_n)$$

$$= \arg \max_{v_j \in V} \frac{P(a_1, a_2, \dots, a_n | v_j) P(v_j)}{P(a_1, a_2, \dots, a_n)}$$

$$= \arg \max_{v_j \in V} P(a_1, a_2, \dots, a_n | v_j) P(v_j)$$

Bayesの定理
無視しても結構

- 問題: $P(a_1 \dots a_n | v_j)$ を評価するのに大量のデータが必要. 2値属性としても、属性数が n なら $10 \cdot 2^n$ 個程度のサンプルが必要

Naïve であること

- ありえない仮定をおく
- Naïve Bayes の仮定: 属性同士は、属するクラスが所与なら、独立である。すなわち
 - $P(a_1, \dots, a_n | v_j) = P(a_1 | v_j) P(a_2 | v_j) \dots P(a_n | v_j)$
 - 条件付独立性 (クラスが所与の時) とも
- この仮定のもと、 v_{MAP} は

$$v_{NB} = \arg \max_{v_j \in V} P(a_1, a_2, \dots, a_n | v_j) P(v_j)$$

$$= \arg \max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

「ある事例について」を明記すると

- ある事例 E の属するクラス (事例に未知の属性があつて、それが「クラス」であつてよい) を推定する

$$v_{MAP} = \arg \max_{v_j \in V} P(v_j | a_1, a_2, \dots, a_n, E)$$

$$= \arg \max_{v_j \in V} \frac{P(a_1, a_2, \dots, a_n | v_j, E) P(v_j | E)}{P(a_1, a_2, \dots, a_n | E)}$$

$$= \arg \max_{v_j \in V} P(a_1, a_2, \dots, a_n | v_j, E) P(v_j | E)$$

- naïve Bayes であれば

$$v_{NB} = \arg \max_{v_j \in V} P(v_j | a_1, a_2, \dots, a_n, E)$$

$$= \arg \max_{v_j \in V} P(a_1, a_2, \dots, a_n | v_j, E) P(v_j | E)$$

$$= \arg \max_{v_j \in V} P(v_j | E) \prod_i P(a_i | v_j, E)$$

蛇足: naïve とは

英辞郎より

「ナイーブ」とはかなり意味が違う。ご注意。

- 【名】うぶな人
- 【形】
- 世間知らずの、[世間を]なめた、経験の少ない、[思考が]単純な、たまされやすい、ばか正直な、うぶな、無警戒な、認識の甘い、愚直な
 - How can you be so naive?: よくそんなに単純でいられるなー。
 - In spite of her appearance, she has a naive side.: ああ見えても彼女はおうぶなところがある。
 - Society is not so naive.: 世間はそれほど甘いものではない。
 - You're naive!: まだ青いな!
 - You're too naive.: おまえは世間知らずだ。
- 純真な、純情な、天真らんまん、あどけない、素朴な、純朴な、無邪気な、初々しい
- 批判する能力に欠ける
- [マウスなどが]実験未使用の、投薬を受けたことがない

解説が必要ですね

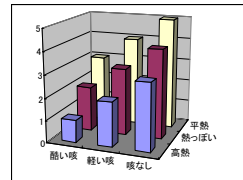
- まず、何をすべきであったか?
- すべきことは、 $p(m)$ と $p(x|m)$ の推定。
- どうすればできるか?
- 属性もクラスも離散変数であるとしよう。
- $p(m)$ は、例えば、 $p(\text{風邪})$ と $p(\neg\text{風邪})$ であるが、これは、数を数えれば、推定できる。
 - 統計を取ればよい。 $p(\text{風邪})$ は「全患者数分の風邪の患者数」で推定できる。
- $p(x|m)$ も同様である。数を数えればよい。
 - $X = \langle \text{咳}, \text{熱} \rangle$ であるとするれば、例えば、 $p(\text{酷い咳}, \text{高熱風邪})$ は、「風邪の患者数分の酷い咳・高熱がある患者数」で推定できる。

解説が必要ですね

- 次に、確率変数間の独立性。
- 二つの確率変数 X, Y 間では常に $P(X, Y) = P(X)P(Y)$ となるなら、この二変数は独立であるという
- X と Y の間に因果関係や相関関係があると独立ではなくなる。
- 独立なら、因果関係や相関関係がないということ。
- しかし、ここで (naïve Bayes で) 重要なことは、確率の計算が簡単になることであって、理論的に or 実質的に独立かどうかではない。
- どうして簡単になるかって?

独立でないとなが大変か

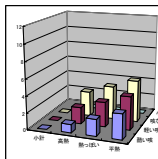
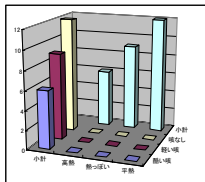
- X の値は、酷い咳、軽い咳、咳なし、 Y の値は高熱、熱っぽい、平熱、とする。
- まずは、独立でないときの話し。
- モデル(分布)を推定するには、これら 3×3 個の組合せについて、頻度を推定する必要がある。
 - 数を数えればよい
- これらの組合せは排他的だから、一つのサンプルは一箇所にしか入れない。
- 頻度を推定するには、(確率0を除き)各組合せに最低でも1個のサンプルが必要(勿論、本当に1個しかないのではお話にならないが)。
- つまり、3値の属性が n 個あれば 3^n 個、2値の属性が n 個あれば 2^n 個のサンプルは、絶対に必要(それでも不十分)ということになる。
- 係数を気にしなくてすむようにオーダー記法を用いれば、 $O(3^n)$ 個や $O(2^n)$ 個 必要、ということになる



この図(いい加減な図)では、9個の属性の組み合わせに対し、27個のサンプルとなっている

独立であると何がよいか

- 独立であるとしよう。
- $p(X, Y) = p(X)p(Y)$ であるから、 3×3 個の $p(X, Y)$ を推定するのに、実は、3個の $p(X)$ と3個の $p(Y)$ を推定すればよいことになる
 - 周辺分布を推定すればよいということ
- もともと、 3×3 個の組合せについて頻度を推定する必要があるのだが、 $3 + 3$ 個に対して推定すればよいことになる。
- つまり、絶対に必要なサンプル数が、3値の属性が n 個あれば $3n$ 個、2値の属性が n 個あれば $2n$ 個に減る。
- つまり、必要なサンプル数は、オーダー記法を用いれば、 $O(n)$ 個ということになる。
- なお、排他性も弱くなっている。属性一つ一つの中では排他性がある(一つのサンプルは一つの属性値しかとれない)が、複数個の属性に渡る排他性はない。右図で、「熱」の中、「咳」の中では排他的だが、それだけ、ということ。



独立ならよいことばかりか?

- 本当に独立なら、よいことばかりである。
- しかし、本当に独立なわけではない
 - 咳が酷ければ、喉が炎症を起こし、熱が出る
- 独立でないのに、独立を仮定すると何が起こるか?
- めちゃくちゃになる(何をしているか分からなくなる)はず。
 - 実際、naïve Bayesによって推定した確率値はまったく合っていないといわれている。
- しかし、実際には、naïve Bayes がうまく機能することが多い。これは、
 - 誤った独立性仮定による誤りの増加より、独立性仮定によってパラメータ数を減らしてパラメータの推定精度を向上させたことによる誤りの減少が勝っている。
 - 分布を推定しているわけではなく、クラス・分類を推定しているのである。
 - 実際には、独立でなくとも独立として十分近似できることが多いからではないかと考えられる。

多項分布

- 実は、これまで曖昧にしてきたのだが、確率分布として、多項分布を仮定していた。
 - 確率変数が離散値をとるときには、よく使用する
- 多項分布は
 - n 個の独立なベルヌーイ試行の結果の分布であり
 - 各試行の結果は有限個 (k 個) の値をとり、
 - それぞれの値をとる確率は p_1, \dots, p_k
 - 確率変数 X_i は n 回の試行で i 番目の結果が起こる回数を示すとするときの確率分布であり、
 - それは

$$p(x_1, \dots, x_k; n, p_1, \dots, p_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}$$

となる

多項分布: naïve Bayes の場合

- naïve Bayes の場合、各属性値は、各インスタンス (各人) について、そうなるかならないか (高熱か否か) を表していることが多い。
- 前のスライドでいえば、 $x_i = 1$ or 0 というような分布である。
- 例えば、 p_1 = 高熱となる確率、 p_2 = 熱っぽい確率、 p_3 = 平熱の確率、とすれば、ある人について

$$p(\text{高熱}) = \prod_{i=1}^n p_1^{x_i} p_2^{0-x_i} p_3^{0-x_i} = p_1, p(\text{熱っぽい}) = \prod_{i=1}^n p_1^{0-x_i} p_2^{x_i} p_3^{0-x_i} = p_2, p(\text{平熱}) = \prod_{i=1}^n p_1^{0-x_i} p_2^{0-x_i} p_3^{x_i} = p_3$$

- これを n 人の風邪の患者 (n 個の事例 (風邪患者)) について調べたところ、 x_1 人が高熱、 x_2 人が熱っぽく、 x_3 人が平熱であったとしよう。n 人の風邪患者についてそうなる確率は、

$$p(x_1, x_2, x_3; n, p_1, p_2, p_3) = \frac{n!}{x_1! x_2! x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}$$

多項分布: パラメータ推定

- これを n 人の風邪の患者 (n 個の事例 (風邪患者)) について調べたところ、 n_1 人が高熱、 n_2 人が熱っぽく、 n_3 人が平熱であったとしよう。n 人の風邪患者についてそうなる確率は、

$$p(n_1, n_2, n_3; n, p_1, p_2, p_3) = \frac{n!}{n_1! n_2! n_3!} p_1^{n_1} p_2^{n_2} p_3^{n_3}$$

- n_1, n_2, n_3 が与えられたとき、これを最大にする p_1, p_2, p_3 は

$$\hat{p}_1 = \frac{n_1}{n}, \hat{p}_2 = \frac{n_2}{n}, \hat{p}_3 = \frac{n_3}{n}$$

- すなわち、 $\left(\frac{n_1}{n}, \frac{n_2}{n}, \frac{n_3}{n}\right)$ は (p_1, p_2, p_3) の最尤推定量である

本来は最初のスライド: Naïve Bayes 分類方法

- 単純だが (だから?) よく知られた方法
 - 単純な割には高精度
 - 単純なだけに、高速
- Bayes 推論を用いて分類クラスを決定する分類課題において、属性値の (分類クラスを条件とする) 条件付独立性を仮定した方法。
- Bayes 定理 + 仮定 **条件付独立**
 - 実際には成り立たないことが多い仮定
 - それに関わらず、実際にはしばしば、良い結果が得られる
- 成功事例:
 - 文書分類
 - その他、なんでも。まずはこれを試す

簡単な例で: 天気とテニス

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

未知の一日

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

この例では、結果があるというより、(属性 Play の値が不明な) ある事例があるというべきであろう

Tom Mitchell の Machine Learning という書籍から、よく使われます

天気とテニス

$$v_{NB} = \underset{v_i \in \mathcal{C}}{\operatorname{argmax}} P(v_i | a_1, a_2, \dots, a_n, E) \\ = \underset{v_i \in \mathcal{C}}{\operatorname{argmax}} P(a_1, a_2, \dots, a_n | v_i, E) P(v_i | E) \\ = \underset{v_i \in \mathcal{C}}{\operatorname{argmax}} P(v_i | E) \prod_i P(a_i | v_i, E)$$

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← テストサンプル E

Outlook=Sunny, ..., Windy=True を A と略記
E が右表と同じ情報源から得られたとすると

$$P(\text{Play}=\text{yes} | A, E) \\ = P(A | \text{Play}=\text{yes}, E) * P(\text{Play}=\text{yes} | E) \\ = P(\text{Outlook}=\text{Sunny} | \text{Play}=\text{yes}, E) \\ * P(\text{Temp}=\text{Cool} | \text{Play}=\text{yes}, E) \\ * P(\text{Humidity}=\text{High} | \text{Play}=\text{yes}, E) \\ * P(\text{Windy}=\text{True} | \text{Play}=\text{yes}, E) \\ * P(\text{Play}=\text{yes}, E) / P(E)$$

さて、 $P(\text{Outlook}=\text{Sunny} | \text{Play}=\text{yes}, E)$ 等はどう考えたらよいのか?

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Overcast	Hot	High	True	No
Rainy	Mild	High	False	Yes
Rainy	Mild	High	True	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

天気とテニス: パラメータ推定

- n回の事例の Outlook について調べたところ、 n_1 回が sunny、 n_2 回が overcast、 n_3 回が rainy であったとしよう。そのなる確率は、 $\frac{n!}{n_1!n_2!n_3!} p_1^{n_1} p_2^{n_2} p_3^{n_3}$
- ただし、 $p_1 = P(\text{Outlook=sunny} \mid \text{右表})$,
 $p_2 = P(\text{Outlook=overcast} \mid \text{右表})$,
 $p_3 = P(\text{Outlook=rainy} \mid \text{右表})$
- p_1, p_2, p_3 の最尤推定量は、 $n_1/n, n_2/n, n_3/n$ となる

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

各確率(の推定値)を計算しよう

	Outlook		Temperature		Humidity		Windy		Play				
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No			
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

天気とテニス

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

テストサンプル E

Outlook=Sunny, ..., Windy=True を A と略記
E が右表と同じ情報源から得られたとすると

$$\begin{aligned}
 &P(\text{Play=yes} \mid A, E) \\
 &= P(A \mid \text{Play=yes}, E) * P(\text{Play=yes} \mid E) \\
 &= P(\text{Outlook=Sunny} \mid \text{Play=yes}, E) \\
 &\quad * P(\text{Temp=Cool} \mid \text{Play=yes}, E) \\
 &\quad * P(\text{Humidity=High} \mid \text{Play=yes}, E) \\
 &\quad * P(\text{Windy=True} \mid \text{Play=yes}, E) \\
 &\quad * P(\text{Play=yes}, E) / P(E) \\
 &= (2/9) * (3/9) * (3/9) * (3/9) * (9/14) / P(E) \\
 &= 0.0053 / P(E)
 \end{aligned}$$

1/P(E) は気にしなくてよい; 比較すべき相手すべてに共通だから。

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

天気とテニス

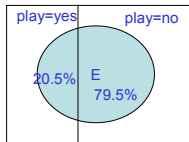
Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

テストサンプル E

$$\begin{aligned}
 &P(\text{Play=yes} \mid A, E) \\
 &= P(A \mid \text{Play=no}, E) * P(\text{Play=no} \mid E) \\
 &= P(\text{Outlook=Sunny} \mid \text{Play=no}, E) \\
 &\quad * P(\text{Temp=Cool} \mid \text{Play=no}, E) \\
 &\quad * P(\text{Humidity=High} \mid \text{Play=no}, E) \\
 &\quad * P(\text{Windy=True} \mid \text{Play=no}, E) \\
 &\quad * P(\text{Play=no}, E) / P(E) \\
 &= (3/5) * (1/5) * (4/5) * (3/5) * (5/14) / P(E) \\
 &= 0.0206 / P(E)
 \end{aligned}$$

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

“気にしなくてよい” 項 P(E) は



$$\begin{aligned}
 &P(\text{Play=yes}, E) + P(\text{Play=no}, E) = P(E) \quad \text{i.e.} \\
 &P(\text{Play=yes}, E)/P(E) + P(\text{Play=no}, E)/P(E) = 1 \quad \text{i.e.} \\
 &P(\text{Play=yes} \mid E) + P(\text{Play=no} \mid E) = 1 \quad \text{i.e.} \\
 &0.0053 / P(E) + 0.0206 / P(E) = 1 \quad \text{i.e.}
 \end{aligned}$$

$$P(E) = 0.0053 + 0.0206$$

というわけで、

$$P(\text{Play=yes} \mid E) = 0.0053 / (0.0053 + 0.0206) = \mathbf{20.5\%}$$

$$P(\text{Play=no} \mid E) = 0.0206 / (0.0053 + 0.0206) = \mathbf{79.5\%}$$

頻度=0 問題

- もしあるクラス値に対してある属性値が一度も起こらなかったらどうなるか (e.g. “Play=Yes” のとき “Humidity = High”)?
 - 確率 $P(\text{Humidity=High} \mid \text{Play=yes})$ は 0 !!
- 従って、事後確率は 0 !!
 - 他の値がどんなに “これは起こりやすい” と言っているても、だ。
- 治療方法:
 - すべてのクラス値-属性値の組に頻度 1 を加える (Laplace correction という);
 - 分母 (正確ではない、右を参照)には k (可能な属性値の個数) を加える (勿論、これと合せて Laplace correction という).

$$\begin{aligned}
 &P(\text{Play=yes} \mid E) \\
 &= P(\text{Outlook=Sunny} \mid \text{Play=yes}) * \\
 &\quad P(\text{Temp=Cool} \mid \text{Play=yes}) * \\
 &\quad P(\text{Humidity=High} \mid \text{Play=yes}) * \\
 &\quad P(\text{Windy=True} \mid \text{Play=yes}) * \\
 &\quad P(\text{play=yes}) / P(E) \\
 &= (2/9) * (3/9) * (3/9) * (3/9) * (9/14) / P(E) \\
 &= 0.0053 / P(E)
 \end{aligned}$$

ではなく:

$$\begin{aligned}
 &\text{‘Outlook’ のとりうる値の個数} \\
 &= ((2+1)/(9+3)) * ((3+1)/(9+3)) * \\
 &\quad ((3+1)/(9+2)) * ((3+1)/(9+2)) * (9/14) / P(E) \\
 &= 0.007 / P(E)
 \end{aligned}$$

‘Windy’ のとりうる値の個数

欠測値問題: 属性値が不明

- 問題: 属性 A の値がない事例があるとどうなるか?
 - しばしば、訓練時やテスト時に、必ずしも全ての属性値が入手できるとは限らない
 - 例: 医療診断
 - <Fever = true, Blood-Pressure = normal, ..., Blood-Test = ?, ...>
 - 値は、本当になかったり、またあっても信頼度が低かったりする
 - 欠測値: 訓練時 versus 分類時
 - 学習時: その属性値のみ無視をする(数えない)
 - 分類時: その属性値の確率は参入しない(1と数える)

欠測値問題

- 一つの単純だが乱暴な解決案:
- 学習時: 当のクラス値-属性値は頻度計算には算入しない
- 分類時: 確率から当の属性は省く
- 例:

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

$$\begin{aligned}
 P(\text{Play}=\text{yes} | E) &= P(\text{Temp}=\text{Cool} | \text{Play}=\text{yes}) * \\
 &= (3/9) * (3/9) * (3/9) * (9/14) / P(E) = 0.0238 / P(E) \\
 P(\text{Play}=\text{no} | E) &= P(\text{Temp}=\text{Cool} | \text{Play}=\text{no}) * \\
 &= (1/5) * (4/5) * (3/5) * (5/14) / P(E) = 0.0343 / P(E) \\
 P(E) \text{を求めれば: } P(\text{Play}=\text{yes} | E) &= 41\%, \quad P(\text{Play}=\text{no} | E) = 59\%
 \end{aligned}$$

数値属性の取扱い

- よくある仮定: 属性値は正規分布をなす(クラス値が所与)
- 正規分布(ガウス分布)のパラメータは2個. 推定値は:

- 標本平均 μ

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- 不偏分散 σ^2

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- 密度関数 $f(x)$:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

天気とテニスの例(改変)

	Outlook		Temperature		Humidity		Windy		Play				
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No			
Sunny	2	3	83	85	86	85	False	6	2	9	5		
Overcast	4	0	70	80	96	90	True	3	3				
Rainy	3	2	68	65	80	70							
						
Sunny	2/9	3/5	mean	73	74.6	mean	79.1	86.2	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	std dev	6.2	7.9	std dev	10.2	9.7	True	3/9	3/5		
Rainy	3/9	2/5											

確率値

$$f(\text{temperature} = 66 | \text{yes}) = \frac{1}{\sqrt{2\pi} \cdot 6.2} e^{-\frac{(66-73)^2}{2 \cdot 6.2^2}} = 0.0340$$

新しい日の分類

- 新たな日 E:

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	True	?

$$\begin{aligned}
 P(\text{Play}=\text{yes} | E) &= P(\text{Outlook}=\text{Sunny} | \text{Play}=\text{yes}) * \\
 &= (2/9) * (0.0340) * (0.0221) * (3/9) * (9/14) / P(E) = 0.000036 / P(E) \\
 P(\text{Play}=\text{no} | E) &= P(\text{Outlook}=\text{Sunny} | \text{Play}=\text{no}) * \\
 &= (3/5) * (0.0291) * (0.0380) * (3/5) * (5/14) / P(E) = 0.000136 / P(E) \\
 P(E) \text{を求めると: } P(\text{play}=\text{yes} | E) &= 20.9\%, \quad P(\text{play}=\text{no} | E) = 79.1\%
 \end{aligned}$$

デモ applet

- naïve Bayes は実に簡単な分類器である。しかし、結構優秀でもある。下記 applet で他の分類器と比較してみよう
 - もとの分布が正規分布ならうまくいくのは当然ではある
 - 他の手法がどのようなものであるかは、今は、知らなくてよい

<http://www.cs.technion.ac.il/~rani/LocBoost/index.html>