دانشگاه آزاد اسلامی واحد تبریز

نام درس: یادگیری ماشین

بخش: یادگیری بیزین

نام استاد: دکتر مسعود کارگر

مقدمه

- استدلال بیزی روشی بر پایه احتمالات برای استنتاج کردن است.

- اساس این روش بر این اصل استوار است که برای هر کمیتی یک توزیع احتمال وجود دارد که با مشاهده یک داده جدید و استدلال در مورد توزیع احتمال آن می‌توان تصمیمات بهینه‌ای اتخاذ کرد.

اهمیت یادگیری بیزی

- در برخی کاربردها (نظیر دسته‌بندی متن) استفاده از روشهای یادگیری بیزی (نظیر دسته‌بندی‌کننده بیزی ساده) توانسته است راه حلهای عملی مفیدی را ارائه کند. نشان داده شده است که کارائی این روش قابل مقایسه با درخت تصمیم و شبکه عصبی است.
- مطالعه یادگیری بیزی به فهم سایر روشهای یادگیری که بطور مستقیم از احتمالات استفاده نمی‌کنند، کمک می‌کند.

ویژگیهای یادگیری بیزی

- مشاهده هر مثال می‌تواند به صورت جزئی باعث افزایش و یا کاهش احتمال درست بودن یک فرضیه گردد.
- برای بدست آوردن احتمال یک فرضیه می‌توان دانش قبلی را با مثال مشاهده شده ترکیب کرد. این دانش قبلی از دو طریق بدست می‌آید:
  1. احتمال قبلی برای هر فرضیه موجود باشد.
  2. برای داده مشاهده شده توزیع احتمال هر فرضیه ممکن موجود باشد.
- روشهای بیزی فرضیه‌هائی ارائه می‌دهند که قادر به پیش‌بینی احتمالی هستند (مثل بیمار به احتمال 93% بهبود می‌یابد).
- مثالهای جدید را می‌توان با ترکیب وزنی چندین فرضیه دسته‌بندی نمود.
- حتی در مواردی که روشهای بیزی قابل محاسبه نباشند، می‌توان از آنها به عنوان معیاری برای ارزیابی روشهای دیگر استفاده کرد.

مشکلات عملی

- نیاز به دانش اولیه در مورد تعداد زیادی از احتمالات دارد. وقتی که این اطلاعات موجود نباشند اغلب ناگزیر به تخمین زدن آن هستیم. برای این کار از اطلاعات زمینه، داده‌هائیکه قبلا جمع‌آوری شده‌اند، و فرضیاتی در مورد توزیع احتمال استفاده می‌شود.

- محاسبه فرضیات بهینه بیزی بسیار هزینه‌بر است (تعداد فرضیه‌های کاندید خطی است.)

ساختار این فصل

- معرفی تئوری بیز، *ML , MAP*

روشهای یادگیری بیزی شامل:

- *Optimal classifier algorithm,*
- *Naive Bayes learning*
- *Bayesian belief network learning*

- رابطه تئوری بیز و سایر روشهای یادگیری

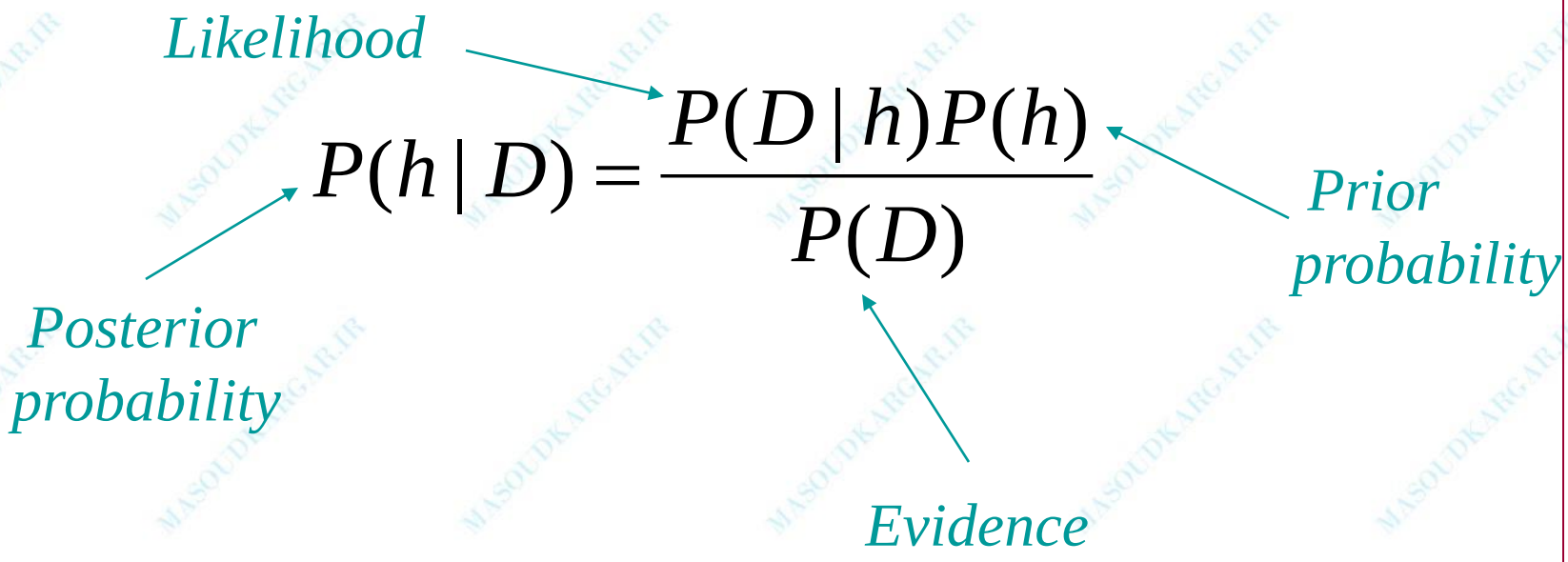
تئوری بیز

- در یادگیری ماشین معمولا در فضای فرضیه H بدنبال **بهترین فرضیه‌ای** هستیم که درمورد داده‌های آموزشی D صدق کند. یک راه تعیین بهترین فرضیه، این است که بدنبال **محتمل‌ترین فرضیه‌ای** باشیم که با داشتن داده‌های آموزشی D و احتمال قبلی در مورد فرضیه‌های مختلف می‌توان انتظار داشت.

- تئوری بیز راه حل مستقیمی در این خصوص ارائه می‌دهد.

تئوری بیز

- سنگ بنای یادگیری بیزی را تئوری بیز تشکیل می‌دهد. این تئوری امکان محاسبه احتمال ثانویه را بر مبنای احتمالات اولیه می‌دهد:



تئوری بیز: تعریف مفاهیم اولیه

- فرض کنید که فضای فرضیه H و مجموعه مثالهای آموزش D موجود باشند. مقادیر احتمال زیر را تعریف می‌کنیم:

1. $P(h)$ = احتمال اولیه‌ای *(prior probablity)* که فرضیه h قبل از مشاهده مثال آموزشی D داشته است. اگر چنین احتمالی موجود نباشد می‌توان به تمامی فرضیه‌ها احتمال یکسانی نسبت داد.
2. $P(D)$ = احتمال اولیه‌ای که داده آموزشی D مشاهده خواهد شد.
3. $P(D|h)$ = احتمال مشاهده داده آموزشی D به فرض آنکه فرضیه h صادق باشد.

- در یادگیری ماشین علاقمند به دانستن $P(h|D)$ یعنی احتمال اینکه با مشاهده داده آموزشی D فرضیه h صادق باشد، هستیم. این رابطه احتمال ثانویه *(posterior probablity)* نامیده می‌شود.
- توجه شود که احتمال اولیه مستقل از داده آموزشی است ولی احتمال ثانویه تاثیر داده آموزشی را منعکس می‌کند.

تئوری بیز

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- همانطور که مشاهده می‌شود با افزایش $P(D)$ مقدار $P(h|D)$ کاهش می‌یابد. زیرا هر چه احتمال مشاهده D مستقل از h بیشتر باشد، به این معنا خواهد بود که D شواهد کمتری در حمایت از h در بر دارد.

Maximum A Posteriori (MAP) hypothesis

- در مسایلی که مجموعه‌ای از فرضیه‌های H وجود داشته و بخواهیم محتمل‌ترین فرضیه را از میان آنان انتخاب کنیم، فرضیه با حداکثر احتمال *Maximum A Posteriori* نامیده می‌شود و از رابطه زیر بدست *(MAP) hypothesis* می‌آید.

$$h_{MAP} \equiv \arg\max_{h \in H} P(h \mid D)$$

$$= \arg\max_{h \in H} \frac{P(D \mid h) P(h)}{P(D)}$$

$$= \arg\max_{h \in H} P(D \mid h) P(h)$$

در این رابطه $P(D)$ یک مقدار ثابت مستقل از h بوده و حذف می‌شود.

Maximum likelihood (ML) hypothesis

- در مواقعی که هیچ اطلاعی در مورد $P(h)$ وجود نداشته باشد می‌توان فرض کرد که تمام فرضیه‌های H دارای احتمال اولیه یکسانی هستند. در این صورت برای محاسبه فرضیه با حداکثر احتمال می‌توان فقط مقدار $P(D|h)$ را در نظر گرفت. این مقدار *likelihood* داده D با فرض h نامیده می‌شود و هر فرضیه‌ای که مقدار آنرا ماکزیمم کند فرضیه *Maximum Likelihood (ML)* نامیده می‌شود:

$$h_{ML} = \arg\max_{h \in H} P(D \mid h)$$

مثال: تشخیص بیماری

- در یک مسئله تشخیص بیماری با دو فرضیه روبرو هستیم:
  1- بیمار دارای سرطان است. *2*- بیمار سالم است.
- داده‌های آزمایشگاهی نشان می‌دهد که *0.008* جمعیت دارای این بیماری هستند.
- به علت نادقیق بودن تست‌های آزمایشگاهی نتایج آن به صورت زیر است:
  - در *98%* مواقعی که شخص واقعا بیمار است نتیجه صحیح مثبت حاصل می‌شود.
  - در *97%* مواقعی که بیمار سالم است نتیجه صحیح منفی حاصل می‌شود.

$$P(cancer)=0.008,\ P(+|cancer)=0.98,\ P(+|\sim cancer)=0.03,$$
$$P(\sim cancer)=0.992,\ P(-|cancer)=0.02,\ P(-|\sim cancer)=0.97$$

مثال: تشخیص بیماری

- حال اگر بیمار جدیدی مشاهده شود که جواب آزمایشگاه مثبت باشد، آیا باید بیمار را مبتلا به سرطان بدانیم؟
- احتمال ابتلای بیمار به سرطان عبارت است از:

$$P(cancer|+) = P(+|cancer)\,P(cancer) / P(+) = (0.98)(0.008) / P(+) = 0.0078 / P(+)$$

- احتمال نداشتن سرطان عبارت است از:

$$P(\sim cancer|+) = P(+|\sim cancer)\,P(\sim cancer) / P(+) = (0.03)(0.992) / P(+) = 0.0298 / P(+)$$

- لذا فرضیه MAP عبارت خواهد بود از:

$$h_{MAP} = \sim cancer$$

مثال: تشخیص بیماری

- $P(cancer|+) + P(\sim cancer|+) = 1$
- $0.0078 / P(+) + 0.0298 / P(+) = 1$
- $P(+) = 0.0078 + 0.0298 = 0.0376$

- احتمال ابتلای بیمار به سرطان عبارت است از:

$$P(cancer|+) = 0.0078 / P(+) = 0.21$$

- احتمال نداشتن سرطان عبارت است از:

$$P(\sim cancer|+) = 0.0298 / P(+) = 0.79$$

خلاصه فرمولهای احتمال

- *Product rule*: probability $P(A \wedge B)$ of a conjunction of two events A and B

$$P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A)$$

- *Sum rule*: probability of a disjunction of two events A and B

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

- *Bayes theorem*: the posterior probability $P(h|D)$ of h given D

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- *Theorem of total probability*: if events $A_1, \ldots, A_n$ are mutually exclusive with $\sum_{i=1}^{n} P(A_i) = 1$, then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Brute-force MAP Learning

- می‌توان با استفاده از تئوری بیزی الگوریتمی برای یادگیری مفهوم ارائه نمود که بتواند فرضیه با بیشترین احتمال را بدست دهد:

Brute-force MAP Learning Algorithm

- برای هر فرضیه h موجود در H مقدار احتمال ثانویه را حساب می‌کنیم.
- فرضیه h_{MAP} را که بیشترین احتمال ثانویه را دارد مشخص می‌کنیم.

دسته‌بندی‌کننده بیزی بهینه
Bayes Optimal Classifier

- الگوریتم *Brute-Force MAP learning* در پی پاسخگوئی به این سوال است: محتمل‌ترین فرضیه برای مجموعه داده آموزشی چیست؟
- در حالیکه اغلب دنبال یافتن پاسخ این سوال هستیم: محتمل‌ترین دسته‌بندی یک نمونه مشاهده شده چیست؟
- اگر چه به نظر می‌رسد که پاسخ سوال دوم را می‌توان با اعمال فرضیه *MAP* به نمونه مورد نظر بدست آورد، روش بهتری برای اینکار وجود دارد:

– در عمل محتمل‌ترین دسته‌بندی برای یک نمونه جدید از ترکیب پیش‌بینی تمامی فرضیه‌ها بدست می‌آید. مقدار پیش‌بینی هر فرضیه در احتمال ثانویه آن ضرب شده و حاصل آنها با هم ترکیب می‌شود.

مثال

- فرض کنید 3 فرضیه $h1, h2, h3$ برای داده‌های آموزشی دارای احتمال ثانویه زیر باشند:

$$P(h1|D) = 0.4, \quad P(h2|D) = 0.3, \quad P(h3|D) = 0.3$$

در نتیجه $h1$ فرضیه MAP می‌باشد.

- اگر به نمونه جدیدی مثل x برخورد کنیم که

$$P(h1) = +, \qquad P(h2) = -, \qquad P(h3) = -$$

- در این صورت احتمال مثبت بودن x برابر با 0.4 و احتمال منفی بودن آن 0.6 است در این صورت دسته‌بندی x چیست؟

دسته‌بندی‌کننده بیزی بهینه
Bayes Optimal Classifier

- در عمل محتمل‌ترین دسته‌بندی برای یک نمونه جدید از ترکیب وزنی پیش‌بینی تمامی فرضیه‌ها بدست می‌آید. اگر دسته‌بندی مثال جدید بتواند هر مقدار v_j از مجموعه V را داشته باشد در این صورت احتمال اینکه مثال جدید دسته‌بندی v_j را داشته باشد برابر است با:

$$P(v_j \mid D) = \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D)$$

- مقدار ماکزیمم رابطه فوق دسته‌بندی بهینه این نمونه را مشخص خواهد نمود:

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D)$$

Bayes Optimal Classification

دسته‌بندی بهینه

Optimal Classification

- برای مثال فوق دسته‌بندی بهینه بیزی به صورت زیر خواهد بود.

$$P(h1|D) = 0.4 \quad P(-|h1) = 0 \quad P(+|h1) = 1$$
$$P(h2|D) = 0.3 \quad P(-|h2) = 1 \quad P(+|h2) = 0$$
$$P(h3|D) = 0.3 \quad P(-|h3) = 1 \quad P(+|h3) = 0$$

- لذا

$$\Sigma i \; P(+|hi)\; P(hi|D) = 0.4 \text{ and}$$
$$\Sigma i \; P(-|hi)\; P(hi|D) = 0.6$$

- در نتیجه این نمونه به صورت منفی دسته‌بندی خواهد شد.

استفاده از این روش برای فضاهای فرضیه‌های بزرگ غیرعملی است.

Naive Bayes Classifier

- یک روش یادگیری بسیار عملی روش *Naive Bayes learner* است. در کاربردهائی نظیر دسته‌بندی متن و تشخیص پزشکی این روش کارائی قابل مقایسه‌ای با شبکه‌های عصبی و درخت تصمیم دارد.
- این روش در مسایلی کاربرد دارد که:
  - نمونه x توسط ترکیب عطفی ویژگیها قابل توصیف بوده و
  - این ویژگیها به صورت شرطی مستقل از یکدیگر باشند.
  - تابع هدف $f(x)$ بتواند هر مقداری را از مجموعه محدود V داشته باشد.
  - مجموعه مثالهای آموزشی نسبتا زیادی در دست باشد.

Naive Bayes Classifier

- تابع هدف زیر را در نظر بگیرید $f : X \rightarrow V$ که در آن هر نمونه x توسط ویژگی زیر مشخص می‌شود: $(a_1,\ldots a_n)$
- صورت مسئله: برای یک نمونه مشاهده شده، مقدار تابع هدف یا به عبارت دیگر دسته‌بندی آنرا مشخص کنید.
- در روش بیزی برای حل مسئله محتمل‌ترین مقدار هدف v_{MAP} محاسبه می‌شود:

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j \mid a_1,\ldots,a_n)$$

این رابطه با استفاده از تئوری بیز به صورت روبرو نوشته می‌شود:

$$\begin{aligned} v_{MAP} &= \arg\max_{v_j \in V} \frac{P(a_1,\ldots,a_n \mid v_j)P(v_j)}{P(a_1,\ldots,a_n)} \\ &= \arg\max_{v_j \in V} P(a_1,\ldots,a_n \mid v_j)P(v_j) \end{aligned}$$

Naive Bayes Classifier

$$v_{MAP} = \arg\max_{v_j \in V} P(a_1, \ldots, a_n \mid v_j) P(v_j)$$

- در رابطه فوق مقدار $P(v_j)$ با شمارش دفعاتی که v_j در مثالهای آموزشی مشاهده شده است محاسبه می‌شود.
- اما محاسبه $P(a_1, \ldots a_n \mid v_j)$ چندان عملی نیست مگر اینکه مجموعه داده آموزشی بسیار بسیار بزرگی در دست باشد.
- روش یادگیری *Naive Bayes Classifier* بر پایه این فرض ساده *(Naive)* عمل می‌کند که:

مقادیر ویژگیها به صورت شرطی مستقل هستند

- در این صورت برای یک مقدار هدف مشخص احتمال مشاهده ترکیب عطفی $(a_1, \ldots a_n)$ برابر است با حاصلضرب احتمال تک تک ویژگیها. در این صورت رابطه فوق به صورت زیر در می‌آید:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_{i=1}^{n} P(a_i \mid v_j)$$

Naive Bayes Classifier

Naive Bayes Classifier

خلاصه:

- در روش یادگیری *Naive Bayes Classifier* مقادیر مختلف $P(v_j)$ و $P(a_i|v_j)$ با استفاده از دفعات تکرار آنها تخمین زده می‌شود.
- مجموعه این تخمین‌ها فرضیه‌ای را تشکیل می‌دهد که با استفاده از رابطه زیر برای دسته‌بندی داده جدید بکار می‌رود:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_{i=1}^{n} P(a_i \mid v_j)$$

- در این روش هیچگونه عمل جستجوی آشکاری در فضای فرضیه وجود ندارد.

مثال

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

- اعمال دسته‌بندی‌کننده ساده بیزی به مسئله دسته‌بندی روزها بر اساس اینکه بازی تنیس در آن انجام خواهد شد یا نه؟
- داده‌های این مثال در جدول زیر نشان داده شده است:

مثال

- می‌خواهیم با این روش دسته مثال زیر را مشخص کنیم:

$$x:(Outl=Sunny,\ Temp=Cool,\ Hum=High, Wind=strong)$$

- با اعمال رابطه دسته‌بندی‌کننده ساده بیزی داریم:

$$v_{NB} = \arg\max_{v_k \in [yes, no]} P(v_k) \prod_i P(a_i \mid v_k)$$

$$= \arg\max_{v_k \in [yes, no]} P(v_k) P(Outlook = sunny \mid v_k) P(Temp = cool \mid v_k)$$

$$P(Humidity = high \mid v_k) P(Wind = strong \mid v_k)$$

- برای محاسبه این مقدار باید تعداد 10 مقدار احتمال را تخمین بزنیم.

مثال

- اینکار با شمارش دفعات تکرار در مثالهای آموزشی انجام می‌شود:

$$P(PlayTennis = yes) = 9/14 = 0.64$$

$$P(PlayTennis = no) = 5/14 = 0.36$$

$$P(Wind = strong \mid PlayTennis = yes) = 3/9 = 0.33$$

$$P(Wind = strong \mid PlayTennis = no) = 3/5 = 0.60$$

etc.

$$P(yes)P(sunny \mid yes)P(cool \mid yes)P(high \mid yes)P(strong \mid yes) = 0.0053$$

$$P(no)P(sunny \mid no)P(cool \mid no)P(high \mid no)P(strong \mid no) = \mathbf{0.0206}$$

$$\Rightarrow answer: PlayTennis(x) = no$$

احتمال مقادیر تخمین

- در مثال قبل مقدار احتمالات بر اساس نسبت تعداد مشاهده شده یک مقدار به تعداد کل حالات ممکن محاسبه گردید n_c/n
- اگرچه این مقدار مشاهده شده می‌تواند تخمین خوبی از مقدار احتمال باشد، با این وجود اگر مقدار n_c خیلی کوچک و یا صفر باشد می‌تواند منجر به نتایج غلطی شود.
- اگر برای دسته v_j مقدار a_i هرگز مشاهده نشود، مقدار $P(a_i|v_j)=0$ شده و در نتیجه کل حاصل‌ضرب صفر خواهد شد. برای جلوگیری از این مشکل از روشی به نام *m-estimate* استفاده می‌شود.

$$\frac{n_c + mp}{n + m}$$

m-estimate of probablity

- که در آن n_c و n همان مقادیر قبلی بوده و p تخمین اولیه از مقدار احتمالی است که به دنبال آن هستیم و m یک مقدار ثابت و تعداد مثالهای مجازی است. معمولا مقدار p به صورت یکنواخت در نظر گرفته شده و برابر با $p = 1/k$ در نظر گرفته می‌شود که k تعداد مقادیر ممکن برای ویژگیهاست. در حقیقت m یک وزنی برای احتمال p است.

دسته‌بندی متن

مثالهائی از دسته‌بندی متن:

- تعیین مقاله‌های مورد علاقه یک شخص.
- دسته‌بندی صفحات وب بر اساس موضوع.

برای چنین کاربردهائی دسته‌بندی‌کننده ساده بیزی می‌تواند بسیار موثر عمل کند. اگرچه در عمل شرط استقلال ویژگیها برقرار نیست. (مثلا احتمال دیدن کلمه ماشین بعد از کلمه یادگیری زیاد است)

دسته‌بندی متن

در طراحی یک راه حل برای چنین مسئله‌ای با دو نکته مواجه هستیم:

1. تصمیم‌گیری در مورد اینکه یک متن دلخواه را چگونه به صورت مقادیر ویژگی نشان دهیم.
2. تصمیم‌گیری در مورد اینکه مقادیر احتمال مورد نیاز را چگونه تخمین بزنیم.

نشان دادن متن به صورت مقادیر ویژگی

- دو راه برای اینکار امتحان شده است:

1. موقعیت هر کلمه در متن به صورت یک ویژگی در نظر گرفته می‌شود. مثلا متنی که 100 کلمه دارد دارای 100 ویژگی نیز خواهد بود.
2. هر کلمه موجود در فرهنگ لغات به عنوان یک ویژگی در نظر گرفته شده (حدود 50000) و تعداد تکرار آنها در متن شمارش می‌شود.

نمایش متن

- در روش اول هر متن به برداری ازکلمات تبدیل شده و بازای موقعیت هر کلمه یک ویژگی نسبت داده می‌شود. که مقدار آن ویژگی برابر با آن کلمه خواهد بود. $doc = (a_1=w_1, a_i=w_k, \ldots, a_n=w_n)$

- برای مثال فرض کنید که از تعداد *1000* متن آموزشی تعداد *700* متن به صورت *dislike* و *300* متن به صورت *like* دسته‌بندی شده باشند. در این صورت برای دسته‌بندی یک متن جدید با *111* کلمه می‌توان رابطه دسته‌بندی‌کننده ساده بیزی را به صورت زیر بکار برد:

$$v_{NB} = \underset{v_k \in [dislike, like]}{\arg\max} \; P(v_k) \prod_{i=1}^{111} P(a_i \mid v_k)$$

مثال از نمایش متن

Our approach to representing arbitrary text documents is disturbingly simple: Given a text document, such as this paragraph, we define an attribute for each word position in the document and define the value of that attribute to be the English word found in that position. Thus, the current paragraph would be described by 111 attribute values, corresponding to the 111 word positions. The value of the first attribute is the word "our," the value of the second attribute is the word "approach," and so on. Notice that long text documents will require a larger number of attributes than short documents. As we shall see, this will not cause us any trouble.

$$v_{NB} = \arg\max_{v_k \in [dislike, like]} P(v_k) \prod_{i=1}^{111} P(a_i \mid v_k)$$

$$= \arg\max_{v_k \in [dislike, like]} P(v_k) P(a_1 = "Our" \mid v_k) P(a_2 = "approach" \mid v_k)$$

$$\ldots P(a_{111} = "trouble" \mid v_k)$$

یک اشکال اساسی

- استفاده از فرض استقلال بیزی در این مثال به وضوح غلط است یعنی نمی‌توان فرض کرد که احتمال بودن یک کلمه در یک محل مستقل از کلماتی است که در سایر محل‌ها قرار گرفته‌اند. با این وجود چون چاره دیگری نداریم ناگزیر از استفاده از این فرض هستیم. در عمل نشان داده شده که علیرغم این فرض نادرست، استفاده از دسته‌بندی‌کننده بیزی ساده نتایج خوبی داشته است.

محاسبه مقادیر احتمال

- برای محاسبه $P(v_j)$ تعداد هر کلاس در داده آموزشی شمارش می‌شود.
- محاسبه $P(a_i=w_k|v_j)$ (در اینجا w_k لغت kام از فرهنگ لغت است.) بسیار سخت است.
- زیرا باید برای تمام ترکیبات ممکن از موقعیتهای کلمات فرهنگ لغت این احتمال تخمین زده شود. که برای این مثال باید 2*111*50,000 تقریبا معادل با 10,000,000 احتمال از روی مجموعه داده‌های آموزشی تخمین زده شود، که عملی ناشدنی است.
-

محاسبه مقادیر احتمال

- از این رو در عمل فرض می‌شود که احتمال مشاهده یک کلمه مشخص w_k مستقل از محل قرار گرفتن آن باشد.

- به عبارت دیگر کل مجموعه $P(a_1=w_k|v_j)$, $P(a_2=w_k|v_j)$,... با $P(w_k|v_j)$ تخمین زده می‌شود، که در این حال فقط باید $2*50,000$ احتمال محاسبه شود.

- همچنین برای محاسبه مقدار احتمال از روش *m-estimate* استفاده خواهد شد.

$$\frac{n_k+1}{n+|Vocabulary|}$$

الگوریتم یادگیری

LEARN_NAIVE_BAYES_TEXT(*Examples*, *V*)

Examples is a set of text documents along with their target values. V is the set of all possible target values. This function learns the probability terms $P(w_k|v_j)$*, describing the probability that a randomly drawn word from a document in class* v_j *will be the English word* w_k*. It also learns the class prior probabilities* $P(v_j)$*.*

1. *collect all words, punctuation, and other tokens that occur in Examples*
   - *Vocabulary* ← the set of all distinct words and other tokens occurring in any text document from *Examples*
2. *calculate the required* $P(v_j)$ *and* $P(w_k|v_j)$ *probability terms*
   - For each target value v_j in V do
     - $docs_j \leftarrow$ the subset of documents from *Examples* for which the target value is v_j
     - $P(v_j) \leftarrow \frac{|docs_j|}{|Examples|}$
     - $Text_j \leftarrow$ a single document created by concatenating all members of $docs_j$
     - $n \leftarrow$ total number of distinct word positions in $Text_j$
     - for each word w_k in *Vocabulary*
       - $n_k \leftarrow$ number of times word w_k occurs in $Text_j$
       - $P(w_k|v_j) \leftarrow \frac{n_k+1}{n+|Vocabulary|}$

الگوریتم دسته‌بندی

CLASSIFY_NAIVE_BAYES_TEXT (Doc)

Return the estimated target value for the document Doc.

a_i *denotes the word found in the ith position within Doc.*

- *positions* ← *all word positions in Doc that contain tokens found in Vocabulary*
- *Return* v_{NB}, *where*

$$v_{NB} = argmax_{vj\ in\ V} P(v_j) \prod_{i\ in\ positions} P(a_i | v_j)$$

نتایج تجربی: یادگیری گروههای خبری

- هدف: تعین اینکه یک مقاله مورد بررسی به کدام یک از 20 گروه خبری *news group* زیر اختصاص دارد:

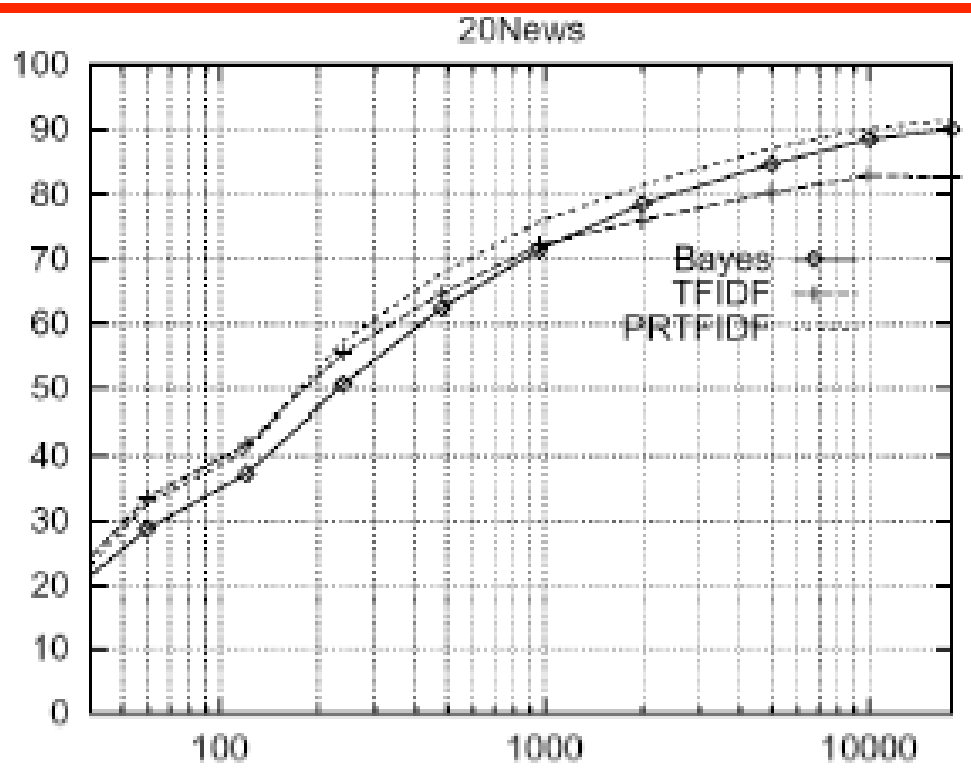
comp.graphics	***misc.forsale***	***alt.atheism***	***soc.religion.christian***
comp.os.ms-windows.misc	***rec.sport.hockey***	***talk.religion.misc***	***sci.space***
comp.windows.x	***rec.autos***	***talk.politics.mideast***	***sci.med***
comp.sys.ibm.pc.hardware	***rec.motorcycles***	***talk.politics.misc***	***sci.crypt***
comp.sys.mac.hardware	***rec.sport.baseball***	***talk.politics.guns***	***sci.electronics***

- داده آموزشی: تعداد 1000 متن به همراه گروه خبری مربوطه.
- نتیجه دسته‌بندی: 89% دقت در دسته‌بندی حاصل گردید.

در این مثال 100 کلمه متداول نظیر the از مجموعه لغات حذف شده است.همجنین کلماتی که تعداد تکرار آنها از 3کمتر بوده نیز حذف گردیده است .در مجموع تعداد 38500 کلمه در لغتنامه وجود داشته است.

منحنی یادگیری

Learning Curve



Accuracy vs. Training set size (1/3 withheld for test)

خلاصه از نگرش بیزی به یادگیری ماشین

نگرش بیزی به یادگیری ماشین (و یا هر فرایند دیگر) به صورت زیر است:

1. دانش موجود در باره موضوع را به صورت احتمالاتی فرموله می‌کنیم.
   - برای اینکار مقادیر کیفی دانش را به صورت توزیع احتمال، فرضیات استقلال و غیره مدل می‌نمائیم. این مدل دارای پارامترهای ناشناخته‌ای خواهد بود.
   - برای هر یک از مقادیر ناشناخته، توزیع احتمال اولیه‌ای در نظر گرفته می‌شود که بازگوکننده باور ما به محتمل بودن هر یک ازاین مقادیر بدون دیدن داده است.
2. داده را جمع‌آوری می‌نمائیم.
3. با مشاهده داده‌ها مقدار توزیع احتمال ثانویه را تخمین می‌زنیم.
4. با استفاده از این احتمال ثانویه:
   - به یک نتیجه‌گیری در مورد عدم قطعیت می‌رسیم.
   - با میانگین‌گیری روی مقادیر احتمال ثانویه پیش‌بینی انجام می‌دهیم.
   - برای کاهش خطای ثانویه مورد انتظار تصمیم‌گیری می‌کنیم.