Bayesian Learning

1400/09/11

Amir Shokri

Bayesian Learning

استدلال بیزین روشی بر پایه احتمالا برای استنتاج کردن است.

اساس کلی بیز این است که برای هر کمیتی یک توزیع احتمال وجود دارد که با مشاهده یک داده جدید و استدلال در مورد توزیع احتمال آن می توان تصمیمات بهینه ای در خصوص آن اتخاذ کرد.

اهمیت یادگیری بیزین :

- کاربرد در دسته بندی متن
- ارائه راه حل های مفید در روش های یادگیری
- کارایی بهتر بیز در مسائل یکسان نسبت به روش های درخت تصمیم و شبکه های عصبی

Bayesian Learning

- دانش موجود درباره ی موضوع را با تعدادی احتمال ذخیره می کنیم.
  - مقادیر کیفی دانش موجود را به صورت توزیع احتمال، فرضیات استقلال و ... مدل می کنیم.
  - مدلی که ایجاد می کنیم دارای پارامترهای ناشناخته خواهد بود.
  - برای هر یک از مقادیر ناشناخته، تزیع احتمال پیشین در نظر گرفته ایم که بازگو کننده ی این احتمال داشتن را در هر یک از مقادیر بدون مشاهده نمونه های داده ایجاد می کند.
- جمع آوری داده ها را انجام می دهیم.
- با مشاهده داده های مختلف مقدار توزیع احتمال پسین را محاسبه می کنیم.
- با احتمال پسین در مورد نمونه های جدید تصمیم گیری های خودمان را انجام می دهیم.

Bayesian Learning

در بیز یادگیری به صورت تدریجی است.

هر نمونه داده باعث افزایش یا کاهش احتمال درست بودن فرضیه ی ما می شود.

احتمالی که برای یک فرضیه در نظر می گیریم با دانش قبلی و اطلاعات موجودی که در آن نمونه مشاهده شده است تصمیم گیری می شود.

پس احتمال قبلی برای هر فرضیه موجود می باشد.

فرضیه های حاصل از روش های بیزین قادر به پیش بینی احتمالی هستند.

داده های جدید را می توان با ترکیب وزنی چندین فرضیه دسته بندی نمود.

به طور کلی در فضای فرضیه یا H با کمک تئوری بیز به دنبال بهترین فرضیه ای خواهیم بود که در مورد داده های آموزشی یا D صدق پیدا کند.

Bayesian Learning

یک راه تعیین بهترین فرضیه گشتن به دنبال محتمل ترین فرضیه می باشد که با داشتن داده ی آموزشی D و احتمال قبلی در مورد فرضیه های مختلف می توان انتظار داشت.

فضای فرضیه H

مجموعه داده آموزشیD

احتمال پیش بینی که فرضیه h قبل از مشاهده مثال آموزشی D را داشته است. P(h)

احتمال پیشین مشاهده داده های آموزشی P(D)

احتمال مشاهده داده آموزشی D به فرض اینکه فرضیه h صادق باشد P(D|h)

Bayesian Learning

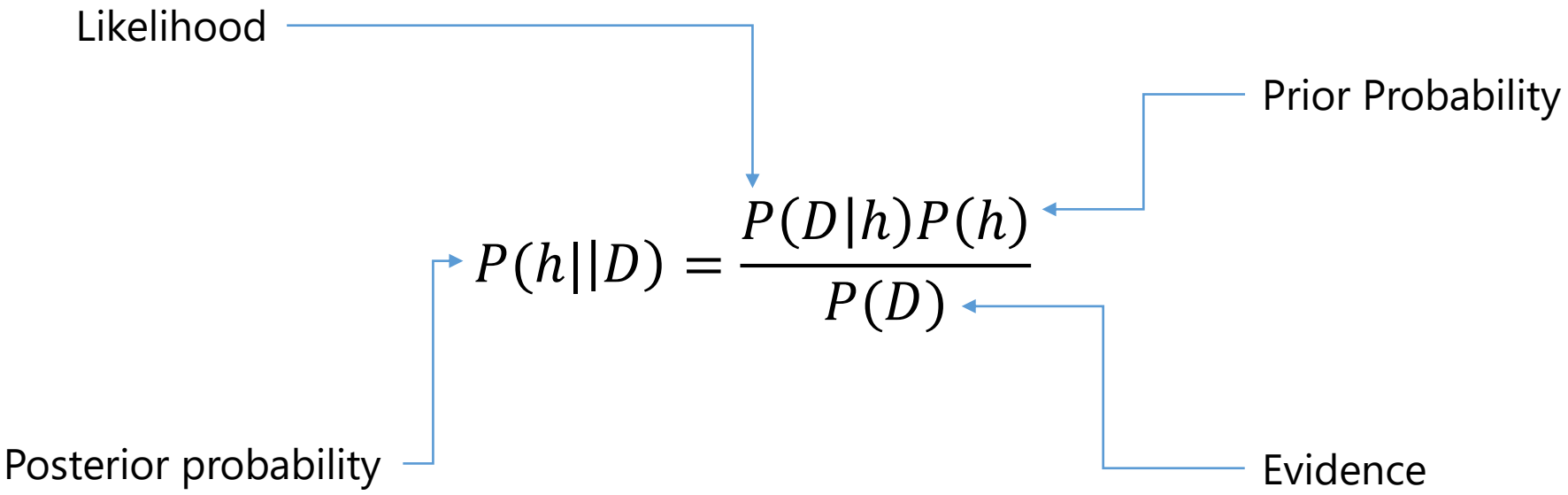
انتظار نهایی ما به دست آمدن $P(h|D)$ است.

مفهوم $P(h|D)$ -> احتمال اینکه با داده ی آموزشی D فرضیه h صادق باشد. (احتمال پسین)

احتمال پیشین همیشه مستقل از D است ولی احتمال پسین متاثر از D است.

یادگیری بیزین بر مبنای نظریه احتمال بیزین است.

Bayesian Learning



Bayesian Learning

نحوه تعیین محتمل ترین فرضیه (فرضیه با حداکثر پسین MAP) از بین مجموعه فرضیات ممکن H :

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

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

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

مستقل از h

Bayesian Learning

در مواقعی که هیچ اطلاعی در مورد P(h) وجود نداشته باشد، می توان فرض کرد که تمام فرضیه های H دارای احتمال پیشین یکسانی هستند.

در اینصورت، برای محاسبه فرضیه با حداکثر احتمال، می توان فقط مقدار P(D|h) را در نظر گرفت. این مقدار، میزان احتمال likelihood داده D نسبت به فرض hنامیده می شود و هر فرضیه ای که این مقدار را بیشینه کند، فرضیه ی Maximum Likelihoodیا ML نامیده می شود:

$$h_{ML} \equiv \arg\max_{h \in H} P(h|D)$$

Bayesian Learning

مثال: تشخیص سرطان

حالات بیمار:

- بیمار دارای بیماری سرطان است.
- بیمار سالم است.

P(cancer)=0.008 P(+|cancer)=0.98 P(+|~cancer)=0.03

P(~cancer)=0.992 P(-|cancer)=0.02 P(-|~cancer)=0.97

Bayesian Learning

مثال: تشخیص سرطان

یک بیمار جدید آمده و نیاز به تصمیم گیری داریم:

احتمال داشتن سرطان

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

احتمال نداشتن سرطان

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

Bayesian Learning

مثال: تشخیص سرطان

فرضیه Map : $h_{map}=\sim cancer$

$P(cancer|+)+P(\sim cancer|+)=1$

$0.0078/P(+)+0.0298/P(+)=1$

$P(+)=0.0078+0.0298=0.0376$

Bayesian Learning

مثال: تشخیص سرطان

احتمال ابتلای بیمار به سرطان:

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

احتمال نداشتن سرطان در بیمار:

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

Bayesian Learning

مفهوم Brute-force Map Learning

می توان با استفاده از نظریه بیزین، الگوریتمی برای یادگیری مفهوم که بتواند فرضیه با بیشترین احتمال را بدست آورد داشته باشیم و آن را Brue-force Map Learning Algorithm بنامیم.

برای هر فرضیه h موجود در H مقدار احتمال پسین را حساب می کنیم.

فرضیه h_{map} را که بیشترین احتمال پسین را دارد مشخص می کنیم.

Bayesian Learning

Bayes Optimal Classifier

دسته بندی بهینه بیزین

Brue-force Map Learning Algorithm : محتمل ترین فرضیه برای مجموعه داده آموزشی داده شده چیست؟

- **محتمل ترین دسته بندی یک نمونه مشاهده شده چیست؟**

در عمل محتمل ترین دسته بندی برای یک نمونه جدید از ترکیب پیش بینی تمامی فرضیه ها بدست می آید. مقدار پیش بینی هر فرضیه در احتمال پسین آن ضرب شده و حاصل آنها با هم ترکیب می شود.

Bayesian Learning

مثال:

ورودی ها:

h1 $P(h_1|D)=0.4$

h2 $P(h_2|D)=0.3$

h3 $P(h_3|D)=0.3$

$P(h_1)=+$ $P(h_2)=-$ $P(h_3)=-$

احتمال مثبت بودن 0.4

احتمال منفی بودن 0.6

Bayesian Learning

دسته بندی بهینه بیزین:

$$P(h_1|D) = 0.4 \; P(-|h_1) = 0 \; P(+|h_1) = 1$$

$$P(h_2|D) = 0.3 \; P(-|h_2) = 1 \; P(+|h_2) = 0$$

$$P(h_3|D) = 0.3 \; P(-|h_3) = 1 \; P(+|h_3) = 0$$

لذا:

$$\sum i \; P(+|h_i)P(h_i|D) = 0.4 \; and \; \sum i \; P(-|h_i)P(h_i|D) = 0.6$$

Bayesian Learning

با توجه به این نمونه های به صورت منفی دسته بندی می شود؛ دقت کنید استفاده از این روش برای فضاهای فرضیه های بزرگ، غیرعملی است.

Naïve Bayes Classifier

کارایی بالا نسبت به شبکه های عصبی و درخت تصمیم

کاربرد:

- نمونه x توسط ترکیب عطفی ویژگی های قابل توصیف باشد.
- این ویژگی ها به صورت شرطی مستقل از یکدیگر باشند.
- تابع هدف f(x) بتواند هر مقداری را از مجموعه محدوده V داشته باشد.
- مجموعه مثال های آموزشی زیاد باشد.

Naïve Bayes Classifier

تابع هدف:

$f:X->V$

$X=(a_1 , ... , a_n)$

صورت مساله : برای یک نمونه مشاهده شده، مقدار تابع هدف یا به عبارت دیگر دسته بندی آن را مشخص کنید.

Naïve Bayes Classifier

حل مسئله روش بیزین برای حل مساله محتمل ترین مقدار هدف v_{MAP} محاسبه می شود:

$$v_{MAP} = \arg\max P(v_j|a_1, \dots, a_n)$$

در این رابطه با استفاده از نظریه بیزین به صورت زیر نوشته می شود:

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

Naïve Bayes Classifier

$$v_{MAP} = \arg\max P(v_j|a_1, \dots, a_n)$$

تعداد دفعاتی که v_j در مثالهای آموزشی مشاهده شده را با $P(v_j)$ می شناسیم و اینکه مصاحبه ی $P(a_1, \dots, a_n|v_j)$ چندان عملی نیست به جز اینکه مجموعه داده آموزشی بسیار بزرگی در دست باشد.

روش یادگیری نایو بیز بر پایه ی فرض ساده (Naive) عمل می کند:

مقادیر ویژگی ها به صورت شرطی مستقل هستند.

برای یک مقدار هدف احتمال مشاهده ی ترکیب عطفی $(a_1, \dots, a_n)$ برابر است با حاصلضرب احتمال تک تک ویژگی ها در اینصورت رابطه ی زیر بدست می آید:

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

Naïve Bayes Classifier

در Naïve Bayes مقادیر مختلف $P(v_j)$ و $P(a_i|v_j)$ با استفاده از تعداد دفعات تکرار تخمین زده می شود.
مجموعه این تخمین ها فرضیه ای را با استفاده از رابطه ی زیر تخمین می زند.

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

در این روش هیچگونه عمل جستجوی آشکاری وجود ندارد.

Bayesian Learning

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	String	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	String	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	strong	no

Amir Shokri

Bayesian Learning

Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	False	N
Sunny	Hot	High	True	N
Overcast	Hot	High	false	P
Rain	Mild	High	False	P
Rain	Cool	Normal	False	P
Rain	Cool	Normal	true	N
Overcast	Cool	Normal	True	P
Sunny	Mild	High	False	N
Sunny	Cool	Normal	False	P
Rain	Mild	Normal	False	P
Sunny	Mild	Normal	True	P
Overcast	Mild	High	True	P
Overcast	Hot	Normal	False	P
Rain	Mild	High	true	N

P(p)=9/14

P(n)=5/14

Bayesian Learning

Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	False	N
Sunny	Hot	High	True	N
Overcast	Hot	High	false	P
Rain	Mild	High	False	P
Rain	Cool	Normal	False	P
Rain	Cool	Normal	true	N
Overcast	Cool	Normal	True	P
Sunny	Mild	High	False	N
Sunny	Cool	Normal	False	P
Rain	Mild	Normal	False	P
Sunny	Mild	Normal	True	P
Overcast	Mild	High	True	P
Overcast	Hot	Normal	False	P
Rain	Mild	High	true	N

Outlook::

P(sunny | p)=2/9
P(sunny | n)=3/5

P(overcast | p)=4/9
P(overcast | n)=0

P(rain | p)=3/9
P(rain | n)=2/5

Bayesian Learning

Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	False	N
Sunny	Hot	High	True	N
Overcast	Hot	High	false	P
Rain	Mild	High	False	P
Rain	Cool	Normal	False	P
Rain	Cool	Normal	true	N
Overcast	Cool	Normal	True	P
Sunny	Mild	High	False	N
Sunny	Cool	Normal	False	P
Rain	Mild	Normal	False	P
Sunny	Mild	Normal	True	P
Overcast	Mild	High	True	P
Overcast	Hot	Normal	False	P
Rain	Mild	High	true	N

temperature::

P(hot | p)=2/9
P(hot | n)=2/5

P(mild| p)=4/9
P(mild | n)=2/5

P(cool | p)=3/9
P(cool | n)=1/5

Bayesian Learning

Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	False	N
Sunny	Hot	High	True	N
Overcast	Hot	High	false	P
Rain	Mild	High	False	P
Rain	Cool	Normal	False	P
Rain	Cool	Normal	true	N
Overcast	Cool	Normal	True	P
Sunny	Mild	High	False	N
Sunny	Cool	Normal	False	P
Rain	Mild	Normal	False	P
Sunny	Mild	Normal	True	P
Overcast	Mild	High	True	P
Overcast	Hot	Normal	False	P
Rain	Mild	High	true	N

humidity::

P(high | p)=3/9
P(high | n)=4/5

P(normal | p)=6/9
P(normal | n)=2/5

Bayesian Learning

Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	False	N
Sunny	Hot	High	True	N
Overcast	Hot	High	false	P
Rain	Mild	High	False	P
Rain	Cool	Normal	False	P
Rain	Cool	Normal	true	N
Overcast	Cool	Normal	True	P
Sunny	Mild	High	False	N
Sunny	Cool	Normal	False	P
Rain	Mild	Normal	False	P
Sunny	Mild	Normal	True	P
Overcast	Mild	High	True	P
Overcast	Hot	Normal	False	P
Rain	Mild	High	true	N

windy::

P(true | p)=3/9
P(true | n)=3/5

P(false | p)=6/9
P(false | n)=2/5

Naïve Bayes Classifier

X: (outlook=sunny, temperature=cool, Humidity=high, wind=strong)

$$V_{NB} = \arg \max_{v_k \in [yes,no]} P(v_k) \prod_{i=1} P(a_i | v_k)$$

$$= \arg \max_{v_k \in [yes,no]} P(v_k) P(outlook = sunny \,| v_k) P(temp$$

Naïve Bayes Classifier

مقدار احتمالات بر اساس نسبت تعداد مشاهده شده یک مقدار به تعداد کل حالات ممکن محاسبه شد: n_c / n

این مقدار می تواند تخمین خوبی از مقدار احتمال باشد با این وجود اگر مقدار n_c خیلی کوچک یا صفر باشد می تواند منجر به نتایج غلط شود.

اگر برای دسته v_j مقدار a_i هرگز مشاهده نشود، مقدار $P(a_i \mid v_j) = 0$ شده و در نتیجه کل حاصلضرب صفر خواهد شد. برای جلوگیری از این مشکل از روش m-estimate استفاده می شود.

$$\frac{n_c + mp}{n + m} \qquad m - estimate\ of\ probability$$

- N و n_c همان مقادیر قبلی است.
- m تعداد مثالهای مجازی است.
- مقدار p هم معمولا $p = 1 / k$ می باشد و به صورت یکنواخت است که k تعداد مقادیر ممکن برای ویژگی هاست.