بسم الله الرحمن الرحیم

هوش مصنوعی پیشرفته

فصل ۲۰

یادگیری مدل‌های احتمالاتی

Learning Probabilistic Models

کاظم فولادی
دانشکده مهندسی برق و کامپیوتر
دانشگاه تهران

http://courses.fouladi.ir/aai

هوش مصنوعی پیشرفته

یادگیری مدل‌های احتمالاتی

۱

یادگیری آماری

یادگیری بیزی

BAYESIAN LEARNING

یادگیری بیزی:

نگاه به **یادگیری** به‌عنوان به‌هنگام‌سازی توزیع احتمال بر روی یک **فضای فرضیه**
با استفاده از قاعده‌ی بیز

یادگیری بیزی

فرمول‌بندی

BAYESIAN LEARNING

متغیر فرضیه H است، با مقادیر
$h_1, h_2, h_3, \ldots$
و احتمال پیشین $\mathbf{P}(H)$

j–امین مشاهده d_j، برآمد متغیر تصادفی D_j را به‌دست می‌دهد.
داده‌های آموزشی:
$\mathbf{d} = d_1, d_2, \ldots, d_N$

با داشتن داده‌ها تاکنون، هر فرضیه یک احتمال پسین دارد:

$$P(h_i|\mathbf{d}) = \alpha P(\mathbf{d}|h_i)P(h_i)$$

درست‌نمایی
Likelihood

پیش‌بینی، از متوسط گیری وزن‌دهی شده با درست‌نمایی‌ها بر روی فرضیه‌ها استفاده می‌کند:

$$\mathbf{P}(X|\mathbf{d}) = \sum_i \mathbf{P}(X|\mathbf{d}, h_i)P(h_i|\mathbf{d}) = \sum_i \mathbf{P}(X|h_i)P(h_i|\mathbf{d})$$

نیازی نیست که یک فرضیه با بهترین حدس را انتخاب کنیم!

یادگیری بیزی

مثال

BAYESIAN LEARNING

فرض می‌کنیم پنج نوع کیسه از آب‌نبات‌ها داریم:

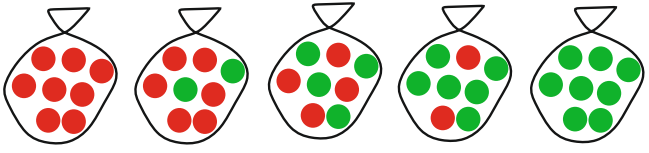
h_1 : 10% آنها 100% آب‌نبات‌ها آلبالویی

h_2 : 20% آنها 75% آب‌نبات‌ها آلبالویی + 25% آب‌نبات‌ها لیمویی

h_3 : 40% آنها 50% آب‌نبات‌ها آلبالویی + 50% آب‌نبات‌ها لیمویی

h_4 : 20% آنها 25% آب‌نبات‌ها آلبالویی + 75% آب‌نبات‌ها لیمویی

h_5 : 10% آنها 100% آب‌نبات‌ها لیمویی



آب‌نبات‌های بیرون کشیده شده از یک کیسه را مشاهده می‌کنیم:

این آب‌نبات‌ها از کدام کیسه بیرون کشیده شده‌اند؟

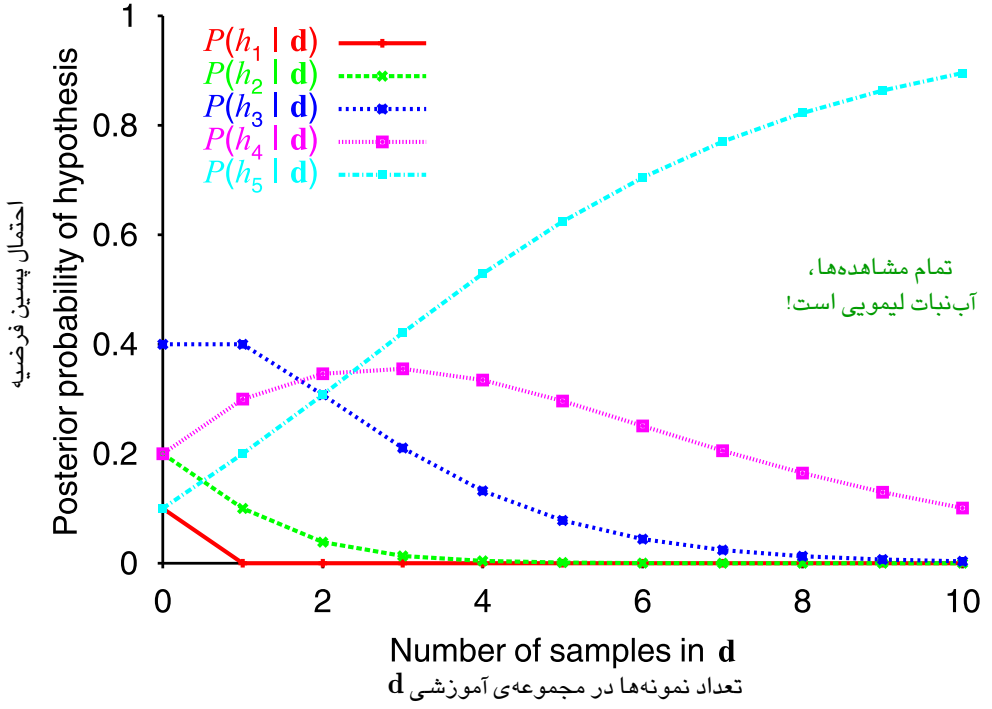
آب‌نبات بعدی دارای چه مزه‌ای است؟

یادگیری بیزی

مثال: احتمال پسین فرضیه‌ها

BAYESIAN LEARNING

$$P(h_i|\mathbf{d}) = \alpha P(\mathbf{d}|h_i)P(h_i)$$

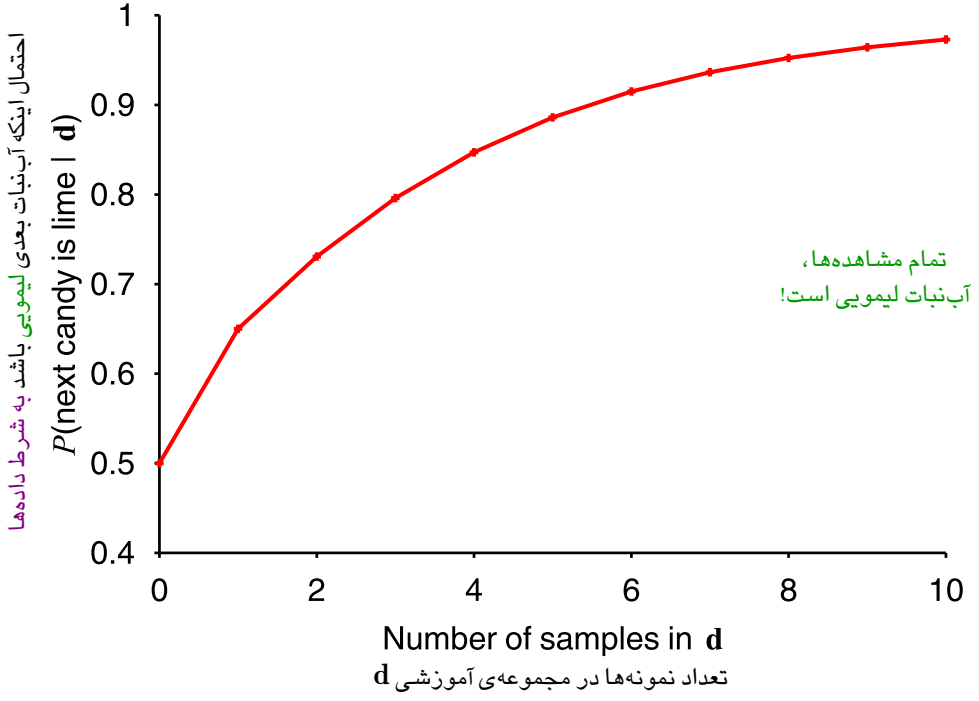


یادگیری بیزی

مثال : احتمال پیش‌بینی

BAYESIAN LEARNING

$$\mathbf{P}(X|\mathbf{d}) = \Sigma_i \, \mathbf{P}(X|\mathbf{d}, h_i)P(h_i|\mathbf{d}) = \Sigma_i \, \mathbf{P}(X|h_i)P(h_i|\mathbf{d})$$



هوش مصنوعی پیشرفته

یادگیری مدل‌های احتمالاتی

۲

یادگیری با داده‌های کامل

تقریب ماکزیمم احتمال پسین

MAXIMUM A POSTERIORI (MAP) APPROXIMATION

مجموع‌یابی بر روی فضای فرضیه‌ها اغلب غیرممکن است.

(برای مثال، 18,446,744,073,709,551,616 تابع بولی با ۶ خصیصه متغیر وجود دارد!)

در عوض، از **یادگیری ماکزیمم احتمال پسین** استفاده می‌کنیم:

Maximum a posteriori (MAP) learning

choose h_{MAP} maximizing $P(h_i|\mathbf{d})$

I.e., maximize $P(\mathbf{d}|h_i)P(h_i)$ or $\log P(\mathbf{d}|h_i) + \log P(h_i)$

برای فرضیه‌های **قطعی**:

$$P(h_i|\mathbf{d}) = \begin{cases} 1 & \text{اگر فرضیه سازگار باشد} \\ 0 & \text{اگر فرضیه ناسازگار باشد} \end{cases}$$

MAP ⇐ ساده‌ترین فرضیه‌ی سازگار

تقریب ماکزیمم احتمال پسین

ایده‌ی پایه‌ی یادگیری توصیف با کمترین طول

THE BASIC IDEA OF MINIMUM DESCRIPTION LENGTH (MDL) LEARNING

Maximum a posteriori (MAP) learning

choose h_{MAP} maximizing $P(h_i|\mathbf{d})$

I.e., maximize $P(\mathbf{d}|h_i)P(h_i)$ or $\log P(\mathbf{d}|h_i) + \log P(h_i)$

جملات لگاریتمی به‌عنوان منفی تعداد بیت‌های لازم برای کدگذاری تعبیر می‌شود.

تعداد بیت لازم برای کدگذاری داده‌ها با داشتن فرضیه

تعداد بیت لازم برای کدگذاری فرضیه

تقریب ماکزیمم درست‌نمایی

MAXIMUM LIKELIHOOD (ML) APPROXIMATION

برای مجموعه‌های داده‌ی بزرگ، احتمال پیشین نامربوط می‌شود

(یعنی می‌توان از آن صرف نظر کرد)

با صرف نظر کردن از احتمال پیشین، به **یادگیری ماکزیمم درست‌نمایی** می‌رسیم:

Maximum likelihood (ML) learning

choose h_{ML} maximizing $P(\mathbf{d}|h_i)$

به طور ساده یعنی: بهترین برازش به داده‌ها را می‌دهد.

معادل با MAP برای احتمال پیشین یکنواخت

(منطقی است، اگر همه‌ی فرضیه‌ها دارای پیچیدگی یکسان باشند: دلیلی بر ترجیح یک فرضیه بر دیگری نداریم)

ML روش استاندارد **یادگیری آماری** (غیر بیزی) است.

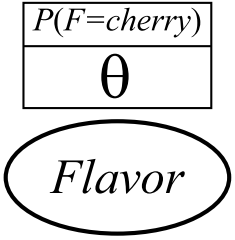
تقریب ماکزیمم درست‌نمایی

یادگیری پارامتر در شبکه‌های بیزی با ماکزیمم درست‌نمایی

ML Parameter Learning in Bayes Nets

یک کیسه از یک کارخانه‌ی جدید رسیده است؛ کسر θ از آب‌نبات‌های آلبالویی؟

- هر θ ممکن است $\Leftarrow$ یک پیوستار از فرضیه‌ها داریم: h_θ
- θ یک پارامتر برای این خانواده‌ی ساده از مدل‌هاست (در اینجا مدل توزیع دوجمله‌ای)



فرض می‌کنیم N آب‌نبات را باز می‌کنیم، c مورد آلبالویی و $\ell = N - c$ مورد لیمویی است.
این مشاهدات مستقل با توزیع یکسان (iid) هستند، پس:

$$P(\mathbf{d}|h_\theta) = \prod_{j=1}^{N} P(d_j|h_\theta) = \theta^c \cdot (1-\theta)^\ell$$

Maximize this w.r.t. θ—which is easier for the log-likelihood:

$$L(\mathbf{d}|h_\theta) = \log P(\mathbf{d}|h_\theta) = \sum_{j=1}^{N} \log P(d_j|h_\theta) = c\log\theta + \ell\log(1-\theta)$$

$$\frac{dL(\mathbf{d}|h_\theta)}{d\theta} = \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 \quad \Rightarrow \quad \theta = \frac{c}{c+\ell} = \frac{c}{N}$$

تقریب ماکزیمم درست‌نمایی

یادگیری چند پارامتر در شبکه‌های بیزی با ماکزیمم درست‌نمایی

ML MULTIPLE PARAMETERS LEARNING IN BAYES NETS

روکش قرمز/سبز به صورت احتمالاتی به مزه‌ی آب‌نبات وابسته است.
درست‌نمایی برای آب‌نبات آلبالویی با روکش سبز می‌شود:

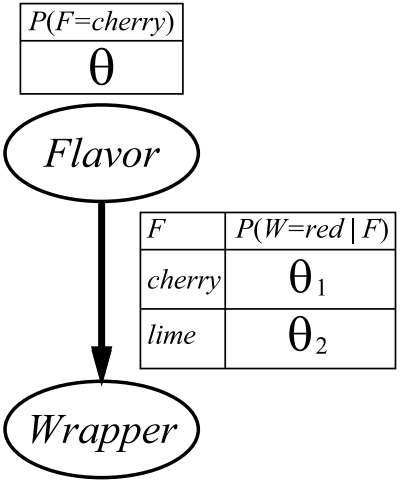
$$P(F = cherry, W = green|h_{\theta,\theta_1,\theta_2})$$
$$= P(F = cherry|h_{\theta,\theta_1,\theta_2})P(W = green|F = cherry, h_{\theta,\theta_1,\theta_2})$$
$$= \theta \cdot (1 - \theta_1)$$

از N آب‌نبات، c مورد آلبالویی و $\ell = N - c$ مورد لیمویی است،

$$P(\mathbf{d}|h_{\theta,\theta_1,\theta_2}) = \theta^c(1-\theta)^\ell \cdot \theta_1^{r_c}(1-\theta_1)^{g_c} \cdot \theta_2^{r_\ell}(1-\theta_2)^{g_\ell}$$

(r_c آب‌نبات آلبالویی با روکش قرمز و g_c با روکش سبز داریم و ...)

$$L = [c\log\theta + \ell\log(1-\theta)]$$
$$+ [r_c\log\theta_1 + g_c\log(1-\theta_1)]$$
$$+ [r_\ell\log\theta_2 + g_\ell\log(1-\theta_2)]$$



تقریب ماکزیمم درست‌نمایی

یادگیری چند پارامتر در شبکه‌های بیزی با ماکزیمم درست‌نمایی: مشتق‌گیری

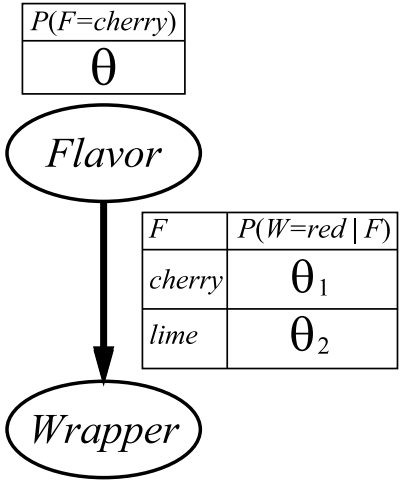
ML MULTIPLE PARAMETERS LEARNING IN BAYES NETS

$$L = [c \log\theta + \ell \log(1-\theta)] + [r_c \log\theta_1 + g_c \log(1-\theta_1)] + [r_\ell \log\theta_2 + g_\ell \log(1-\theta_2)]$$

$$\frac{\partial L}{\partial \theta} = \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 \quad \Rightarrow \quad \theta = \frac{c}{c+\ell}$$

$$\frac{\partial L}{\partial \theta_1} = \frac{r_c}{\theta_1} - \frac{g_c}{1-\theta_1} = 0 \quad \Rightarrow \quad \theta_1 = \frac{r_c}{r_c+g_c}$$

$$\frac{\partial L}{\partial \theta_2} = \frac{r_\ell}{\theta_2} - \frac{g_\ell}{1-\theta_2} = 0 \quad \Rightarrow \quad \theta_2 = \frac{r_\ell}{r_\ell+g_\ell}$$



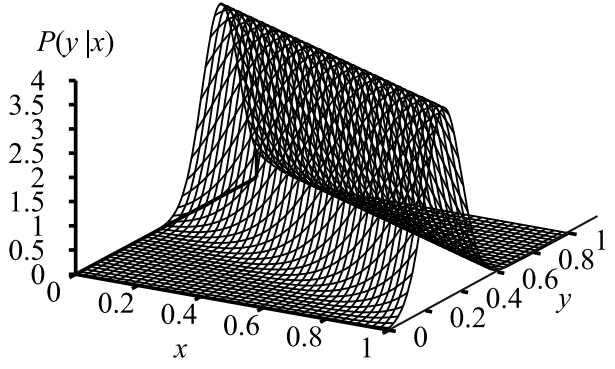
ملاحظه می‌شود که مشتقات L فقط شامل پارامترهای مربوط است.

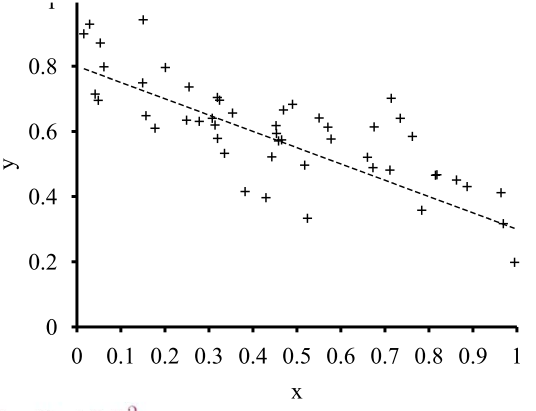
با داده‌های کامل، پارامترها می‌توانند مستقلاً یاد گرفته شوند.

تقریب ماکزیمم درست‌نمایی

مثال: مدل گاوسی خطی

EXAMPLE: LINEAR GAUSSIAN MODEL





$$\text{Maximizing } P(y|x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-(\theta_1 x+\theta_2))^2}{2\sigma^2}} \text{ w.r.t. } \theta_1, \theta_2$$

$$= \text{minimizing } E = \sum_{j=1}^{N} (y_j - (\theta_1 x_j + \theta_2))^2$$

یعنی، می‌نیمم کردن مجموع مربعات خطا، راه‌حل ML را برمی‌گرداند.

برای برازش خطی، نویز گاوسی با واریانس ثابت فرض می‌شود.

تقریب ماکزیمم درست‌نمایی

خلاصه

MAXIMUM LIKELIHOOD (ML) APPROXIMATION

الگوریتم ماکزیمم درست‌نمایی (ML)

یک خانواده‌ی پارامتری از مدل‌ها برای توصیف داده‌ها انتخاب می‌کنیم.
(نیازمند بینش ذاتی و گاهی ارائه‌ی مدل‌های جدید)

↓

درست‌نمایی داده‌ها را به‌صورت تابعی از پارامترها می‌نویسیم.
(ممکن است نیاز به مجموع‌گیری بر روی متغیرهای پنهان داشته باشیم؛ یعنی استنتاج)

↓

مشتق تابع لگاریتم درست‌نمایی را نسبت به پارامترها محاسبه می‌کنیم.

↓

مقادیر پارامترها را با قرار دادن مشتقات مساوی صفر می‌یابیم.
(ممکن است دشوار/غیرممکن باشد؛ می‌توان از تکنیک‌های بهینه‌سازی پیشرفته کمک گرفت.)

یادگیری آماری

خلاصه

STATISTICAL LEARNING

- **یادگیری بیزی** کامل، بهترین پیش‌بینی ممکن را برمی‌گرداند، اما در عمل غیرممکن است (پیچیدگی بالا)!
- یادگیری MAP بین پیچیدگی و دقت بر روی داده‌های آموزشی تعادل برقرار می‌کند.
- یادگیری ML همان MAP است که توزیع پیشین را یکنواخت فرض می‌کند (با فرض مجموعه داده‌ی بزرگ)

هوش مصنوعی پیشرفته

یادگیری مدل‌های احتمالاتی

۳

یادگیری با متغیرهای پنهان: الگوریتم EM

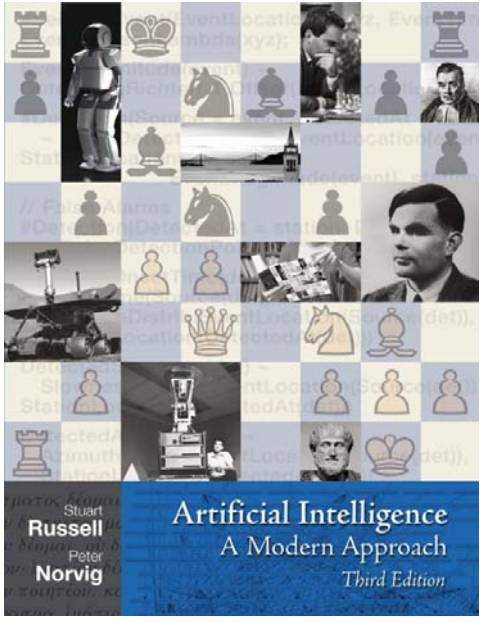
هوش مصنوعی پیشرفته

یادگیری مدل‌های احتمالاتی

۴

منابع، مطالعه، تکلیف

منبع اصلی



Stuart Russell and Peter Norvig,
Artificial Intelligence: A Modern Approach,
3rd Edition, Prentice Hall, 2010.

Chapter 20

20 LEARNING PROBABILISTIC MODELS

In which we view learning as a form of uncertain reasoning from observations.

Chapter 13 pointed out the prevalence of uncertainty in real environments. Agents can handle uncertainty by using the methods of probability and decision theory, but first they must learn their probabilistic theories of the world from experience. This chapter explains how they can do that, by formulating the learning task itself as a process of probabilistic inference (Section 20.1). We will see that a Bayesian view of learning is extremely powerful, providing general solutions to the problems of noise, overfitting, and optimal prediction. It also takes into account the fact that a less-than-omniscient agent can never be certain about which theory of the world is correct, yet must still make decisions by using some theory of the world.

We describe methods for learning probability models—primarily Bayesian networks—in Sections 20.2 and 20.3. Some of the material in this chapter is fairly mathematical, although the general lessons can be understood without plunging into the details. It may benefit the reader to review Chapters 13 and 14 and peek at Appendix A.

20.1 STATISTICAL LEARNING

The key concepts in this chapter, just as in Chapter 18, are **data** and **hypotheses**. Here, the data are **evidence**—that is, instantiations of some or all of the random variables describing the domain. The hypotheses in this chapter are probabilistic theories of how the domain works, including logical theories as a special case.

Consider a simple example. Our favorite Surprise candy comes in two flavors: cherry (yum) and lime (ugh). The manufacturer has a peculiar sense of humor and wraps each piece of candy in the same opaque wrapper, regardless of flavor. The candy is sold in very large bags, of which there are known to be five kinds—again, indistinguishable from the outside:

h_1: 100% cherry,
h_2: 75% cherry + 25% lime,
h_3: 50% cherry + 50% lime,
h_4: 25% cherry + 75% lime,
h_5: 100% lime .

802