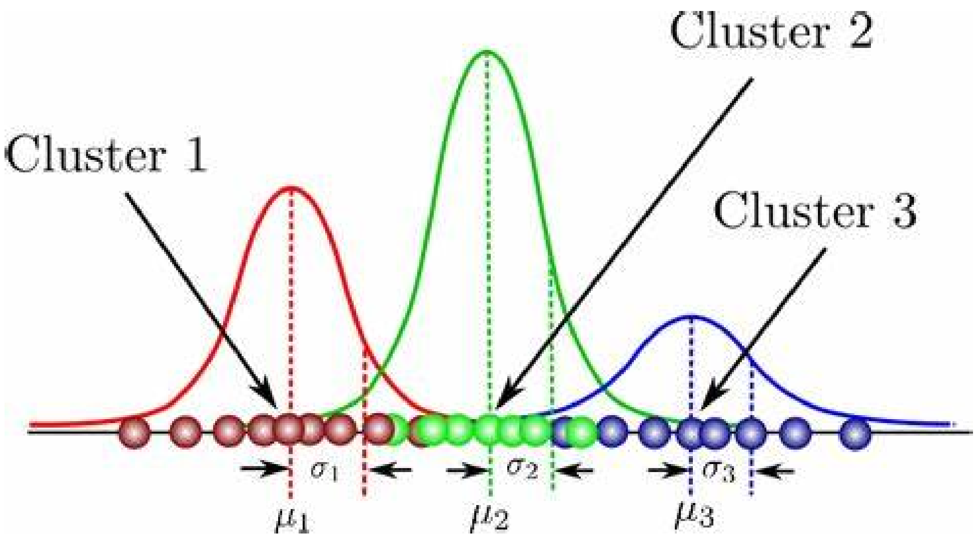
بهنام خدا
تئوری یادگیری ماشین
دکتر سیدصالحی
جلسه چهارم

دانشکده ریاضی و علوم کامپیوتر
اسفند ماه ۱۴۰۲

Probabilistic Perspective of Learning

Probabilistic Unsupervised Learning

یکی از دسته مسائل یادگیری که پیش تر با آن آشنا شده اید مسائل unsupervised هستند. برای این دسته از مسائل نیز میتوان یک دیدگاه احتمالاتی داشت. برای مثال اگر با دیدگاه احتمالاتی به مسئله خوشه بندی نگاه کنید میتوان فرض کرد که هر کدام از خوشه های ما از توزیع های احتمالاتی پیروی میکنند و با استفاده از داده ها به توزیع مدنظر رسید آنگاه هر دیتای جدیدی را میتوان با استفاده از توزیع هایی که بدست آورده ایم به یکی از خوشهها تخصیص داد.



شکل ۱: فرض قرار دادن یک توضیع *gaussian* در دیتا

Generative Approach

دیدگاهی است که در آن فرض میکنیم که در ساختار دادهها توزیع p را داریم و بدنبال یادگرفتن توزیع p میرویم تا بتوانیم کارهایی از قبیل تولید داده انجام دهیم. more information

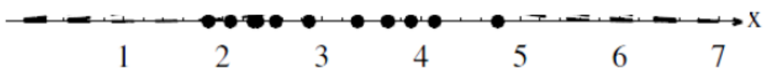
رویکردهای اصلی $density\ estimation$

دو دسته رویکرد برای این نوع تخمین وجود دارد.

- parametric: موردی که فعلا مورد بحث در کورس است و با تخمین پارامترهای توزیع سرکار دارد.
- non-parametric هیچ مدل پارامتری خاصی در نظر گرفته نمی شود.

در parametric ها نیز دو نوع دیدگاه برای تخمین θ وجود دارد بصورتی که دسته ای از افراد یک عدد ثابت را به عنوان پارامتر در نظر میگیرند و حال بدنبال تخمین آن عدد میروند (MLE) درصورتی که دسته دیگر آن را random variable در نظر میگیرند و بدنبال بدست آوردن تخمینی از توزیع آن هستند (MAP, Bayesian Estimator).

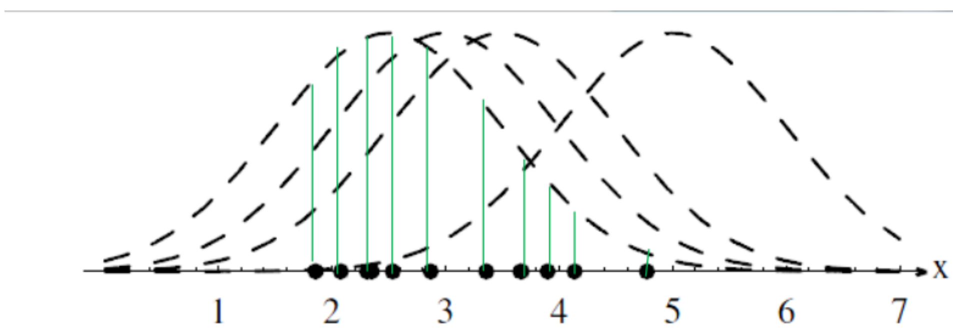
فرض کنید فضای یک بعدی زیر را همراه توضیعی داریم:

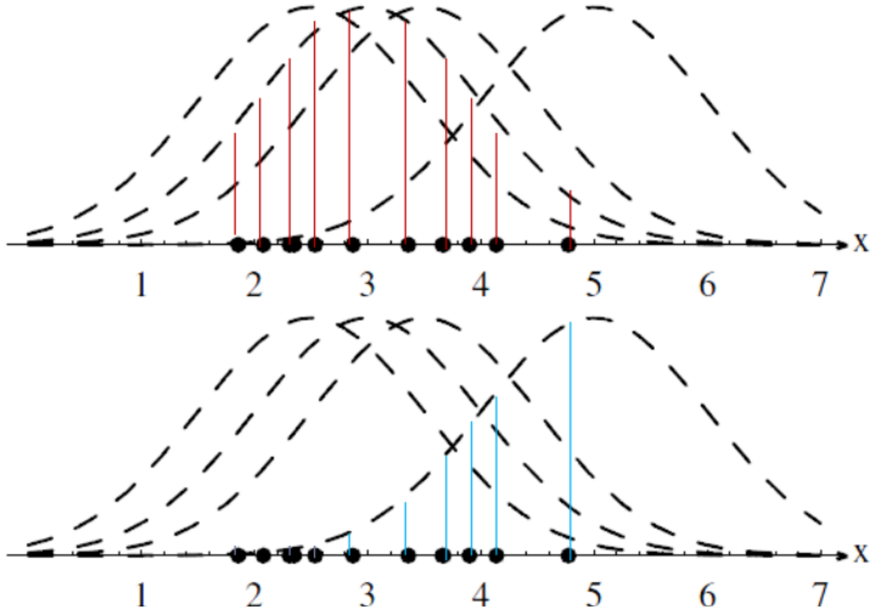


$$P(x|\mu) = N(x|\mu, 1)$$

ما توضیع را نمیدانیم ولی یک خانواده را فرض و میخواهیم از میان آنها توضیعی از میان آن خانواده بخصوص که سمپل ها ازش $generate$ شدن را پیدا کنیم.

خانواده توضیع نرمال را در نظر میگیریم و پارامتر هاش رو تغییر میدیم:





MLE

فرض کنید دسته ای از سمپل ها به شکل $D = \{x^{(1)}, x^{(2)}, ..., x^{(n)}\}$ داریم. همچنین فرض کنید خانواده توزیع نرمال را داشته باشیم در شرایطی که واریانس اعضا ثابت باشد و تنها پارامتری که بین اعضا متفاوت است μ باشد.

حال بدنبال بهترین توزیع نرمالی هستیم که به این داده ها فیت شود. فرض دیگری که میکنیم i.i.d بودن مشاهدات است. پس داریم.

$$p(D|\theta) = \prod_{i=1}^{N} p(x^{(i)}|\theta)$$

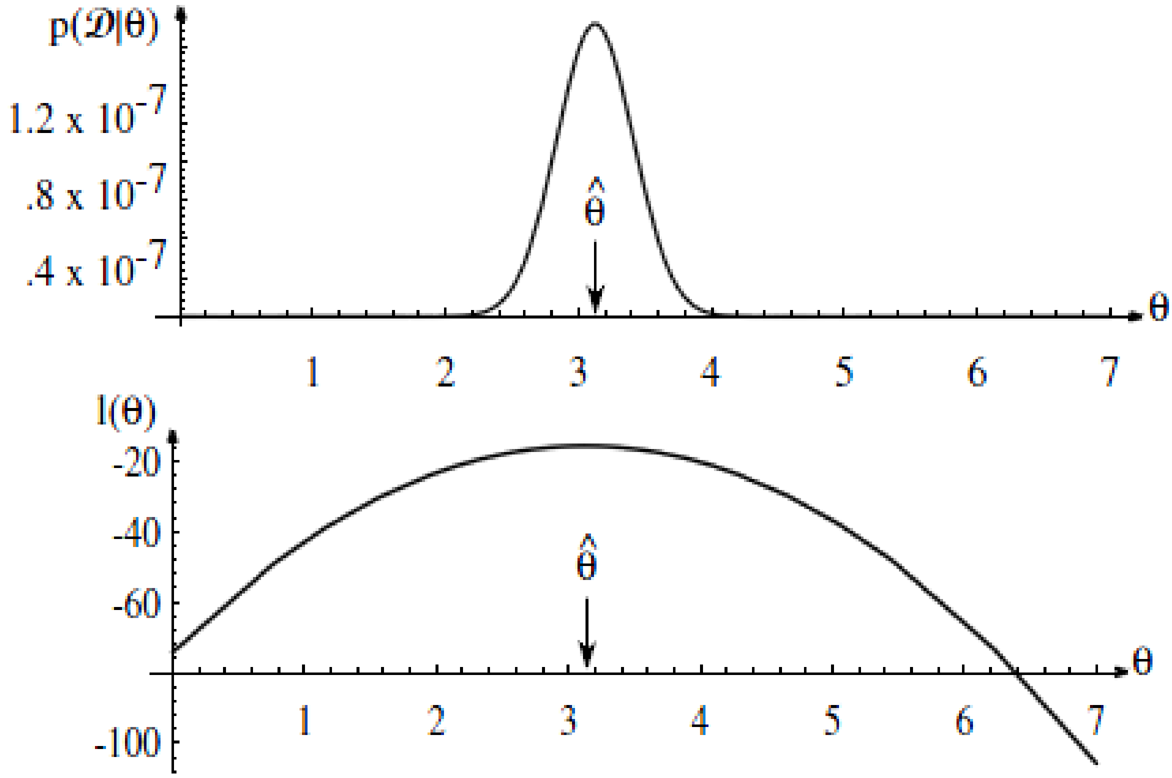
و در MLE بدنبال یافتن مقداری هستیم که این احتمال را بیشینه کند.

$$\hat{\theta}_{ML} = \underset{\theta}{argmax}\, p(D|\theta)$$

برای راحت تر کردن محاسبات هم میتوانیم از فرمت لگاریتمی استفاده کنیم و از انجایی که در اینجا بدنبال argmax گرفتنیم حاصل خود عبارت بالا با لگاریتمش تفاوتی نخواهد داشت.

$$L(\theta) = ln\, p(D|\theta) = ln \prod_{i=1}^{N} p(x^{(i)}|\theta) = \Sigma_{i=1}^{N} ln\, p(x^{(i)}|\theta)$$

$$\hat{\theta}_{ML} = \underset{\theta}{argmax}\, L(\theta) = \underset{\theta}{argmax}\, \Sigma_{i=1}^{N} ln\, p(x^{(i)}|\theta)$$



مثال $Bernoulli$ - گسسته

در این مثال فرض کنید که آزمایش N بار پرتاب یک سکه را داریم بصورتی که m بار سکه شیر آمده است و بقیه دفعات را خط آمده است. پس خانواده توزیع مسئله، توزیع برنولی است که به شکل زیر است.

$$p(x|\theta) = \theta^x(1-\theta)^{1-x}$$

در این مسئله شیر آمدن را ۱ و خط آمدن را ۰ در نظر میگیریم و پارامتر θ را نیز معادل احتمال ۱ شدن یک بار پرتاب سکه میگیریم. بدین ترتیب برای آزمایشمان داریم.

$$p(D|\theta) = \prod_{i=1}^{N} p(x^{(i)}|\theta) = \prod_{i=1}^{N} \theta^{x^{(i)}}(1-\theta)^{1-x^{(i)}}$$

حال از تخمین‌گر استفاده میکنیم.

$$ln\, p(D|\theta) = \Sigma_{i=1}^{N} ln\, p(x^{(i)}|\theta) = \Sigma_{i=1}^{N}\{x^{(i)}\, ln\,\theta + (1-x^{(i)})\, ln\,(1-\theta)\}$$

حال باید مقدار اپتیمال را بدست آوریم.

$$\frac{\partial\, ln\, p(D|\theta)}{\partial\theta} = 0 \rightarrow \theta_{ML} = \frac{\Sigma_{i=1}^{N} x^{(i)}}{N} = \frac{m}{N}$$

ولی این تخمین تخمین مناسبی نیست زیرا حالتی ممکن است پیش بیاد که سمپل ما مثال ۳ از ۳ بار رو آمده باشد و خب با تخمین ۱ میشود که خب مناسب نیست. در اصطلاح به این حالت Overfitting میگویند که بعدا با آشنا میشویم.

۱-۰ مثال $Gaussian$ - پیوسته

این بار بر روی خانواده توزیع نرمال با ثابت در نظر گرفتن پارامتر واریانس بدنبال تخمینی برای μ هستیم.

$$p(x|\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$L(\mu) = ln\, p(x^{(i)}|\mu) = -ln\{\sqrt{2\pi}\sigma\} - \frac{1}{2\sigma^2}(x^{(i)} - \mu)^2$$

و حال برای بدست اوردن مقدار اپتیمال مشتق میگیریم.

$$\frac{\partial L(\mu)}{\partial\mu} = 0 \rightarrow \frac{\partial}{\partial\mu}(\Sigma_{i=1}^{N} ln\, p(x^{(i)}|\mu)) = 0 \rightarrow \Sigma_{i=1}^{N} \frac{1}{\sigma^2}(x^{(i)} - \mu) = 0$$

نتیجتا

$$\hat{\mu}_{ML} = \frac{1}{N}\Sigma_{i=1}^{N} x^{(i)}$$

MAP

در اینجا علاوه بر فرضیاتی که برای تخمین‌گر ML داشتیم فرض میکنیم که پارامتر یک متغیر تصادفی است. همچنین prior distribution آن را نیز داریم. حال بدنبال این هستیم تا پارامتر را به نحوی بدست بیاوریم که توزیع posterior را بیشینه کنیم.

$$\hat{\theta}_{MAP} = \underset{\theta}{argmax}\, p(\theta|D)$$

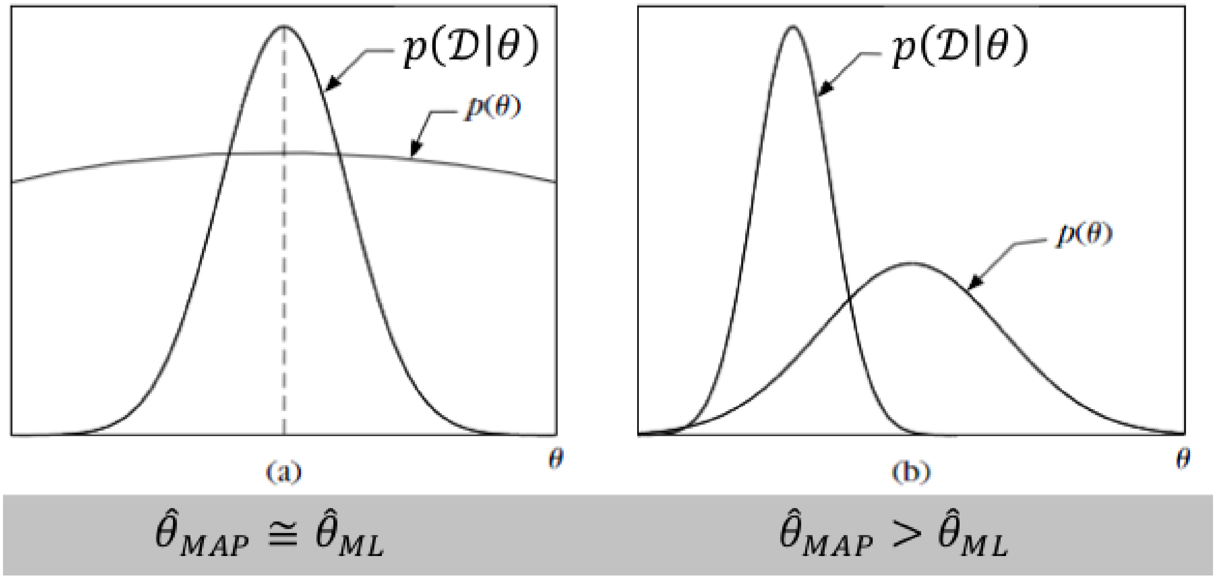
طبق قانون بیز داریم.

$$p(\theta|D) = \frac{p(D|\theta) \times p(\theta)}{p(D)}$$

از انجایی که $p(D)$ وابستگی به پارامتر مدنظر ندارد پس میتوانیم از آن صرفنظر کنیم پس به عبارت زیر میرسیم.

$$\hat{\theta}_{MAP} = \underset{\theta}{argmax}\, p(D|\theta)p(\theta)$$

کاری که در این دیدگاه صورت میگیرد به این ترتیب است که با مشاهده سمپل ها و توزیع انها به تغییر توزیع پیشفرض پارامتر میپردازد و بواسطه دیده هایش توزیع را بروز میکند.
نکته دیگر اینکه لزوما حاصل تخمین MAP و MLE با هم برابر نیستند.



مثال پرتاب سکه

دوباره به این مثال برمیگردیم. همانطور که در بخش قبل داشتیم تابع likelihood برابر است با

$$p(D|\theta) = \prod_{i=1}^{N} \theta^{x^{(i)}}(1-\theta)^{(1-x^{(i)})}$$

و فرض میکنیم که توزیع prior مان نیز از نوع توزیع بتا است.

$$p(\theta) = Beta(\theta|\alpha_1, \alpha_0) = \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$$

بدین ترتیب برای توزیع posterior داریم.

$$= \theta^{m+\alpha_1-1}(1-\theta)^{N-m+\alpha_0-1} \rightarrow p(\theta|D) = Beta(\theta|\alpha_1^{'}, \alpha_0^{'})$$

و

$$\alpha_1' = \alpha_1 + m$$

$$\alpha_0' = \alpha_0 + N - m$$