بازشناسی آماری الگو

جلسه چهارم

یادگیری- تخمین پارامتری تابع توزیع

فهرست مطالب

- یادآوری کلی بحث
- شرایط عملی مسئله کلاسه بندی- در دست بودن مثال ها
- یادگیری و انواع آن
- انواع مسائل و روش های یادگیری در SPR
- تخمین تابع توزیع
  - تخمین پارامتری و غیر پارامتری
- تخمین بیشترین شباهت
- تخمین Bayesian
- مقایسه دو روش
- خلاصه بحث

یادآوری

- در SPR تمام اطلاعات به صورت اطلاعات احتمال از کلاس های مختلف در می آید
  1. احتمال پیشین عضویت در کلاس
  2. توزیع اعضای هر کلاس در فضای مشخصه ها
  3. احتمال مشاهده
  4. احتمال پسین وقوع هر کلاس (در صورت مشاهده نمونه)
- در صورت موجود بودن همه اطلاعات (۱ تا ۳) می توان احتمال پسین (شرطی) وقوع هر کلاس را محاسبه کرد.
- بر اساس تئوری Bayes می توان کلاسی را انتخاب کرد که نمونه مورد بررسی دارای بیشترین احتمال پسین عضویت در آن باشد.

چند مفهوم اساسی

- مفهوم مناطق تصمیم و مرز های تصمیم
- استراتژی های جدا سازی
  - جدا سازی بر اساس بیشترین نمره (مثلا احتمال پسین)
  - جدا سازی بر اساس کمترین فاصله
- روند طراحی- تبدیل مسئله طراحی کلاسیفابر به یک مسئله بهینه سازی:
  - معیار بهینه سازی
    - رویکرد کم ترین خطا
    - رویکرد کم ترین ریسک
  - روش جستجو
- محاسبه احتمال خطا
  - مجموع احتمال وقوع کلاس های غلط در نواحی متعلق به کلاس های درست
  - محاسبه حدود
- قدرت تعمیم
  - خطا روی نمونه هائی که در یادگیری دیده نشده اند

ادامه

- مفهوم کرنل: نگاشت ورودی ها به فضای جدیدی که در آن جدا سازی راحت تر باشد.
- نمونه: توابع جدا ساز Discriminate functions
- کاندیدا های توابع جدا ساز
  - تابع احتمال پسین
  - تابع چگالی احتمال
  - هر تابع یکنوایی که روی تابع چگالی احتمال، اعمال شود(مانند لگاریتم)
    - اگر لگاریتم روی توابع چگالی با توزیع گوسی اعمال شود به توابع جدا ساز خطی منجر می شود.

مقدمه بحث امروز

- آنچه در بحث جلسه گذشته، در طراحی کلاسیفایر ها دیدیم بر مبنای این فرض ها بود که:
  - شیوه بازنمایی نمونه ها مشخص و معلوم است
  - تعداد و چگونگی کلاس ها و نسبت اعضای آنها به هم مشخص است
  - توزیع اعضای هر کلاس معلوم است
- در این شرایط، کلاسیفایر Bayesian پیشنهاد شد.
  - تصمیم گیری بر اساس محاسبه احتمال پسین عضویت یک نمونه در کلاس ها
  - انتساب نمونه به کلاسی که دارای بیشترین احتمال عضویت است
- این شیوه در شکل های مختلف به توابع جدا ساز نیز قابل تبدیل است.

مقدمه بحث امروز..

- از نظر عملی:
  - آنچه در دست داریم، نه توابع توزیع احتمال، بلکه تنها امکان تهیه تعدادی مثال (نمونه) است.
  - در عمل دستیابی به نمونه ها و چگونگی تولید یا دسترسی به آنها نیز در مسائل عملی بسیار تعیین کننده است ولی فعلا در این بحث به آن پرداخته نمی شود.
  - این نمونه ها بایستی به همه پیش فرض های لازم برای استفاده از روش یاد شده در جلسه قبل ، یعنی موجود بودن احتمال پیشین، چگالی احتمال و احتمال مشاهده پاسخ دهد.

شکل های عملی مسئله استفاده از نمونه ها در SPR

- طراحی عملی کلاسیفایرBayesian (انتخاب مرز تصمیم و یا ساختار و پارامتر ها در ...)
  - شرط لازم: معلوم بودن چگالی توزیع و احتمال پیشین
- پیدا کردن چگالی توزیع از روی داده ها (مثلا برای استفاده در فرمول Bayes)
  - تخمین پارامتری
  - تخمین غیر پارامتری
- طراحی کلاسیفایر های غیر از Bayesian با استفاده از نمونه ها
  - نزدیکترین همسایگی، درخت تصمیم، SVM، شبکه های عصبی ....
- بازنمایی ریاضی نمونه ها
  - استخراج و انتخاب مشخصه ها
- محاسبه کارآیی کلاسیفایر

یادگیری از روی مثال ها

- در بحث ما، فرض بر این است که ما سیستمی داریم که در آن تعدادی قدرت انتخاب option و یا پارامتر های آزاد وجود دارد.
- می خواهیم با استفاده از تعدادی مثال، پارامتر های آزاد سیستم را به نحوی انتخاب کنیم(سیستم را به گونه ای تغییر دهیم) که بیشترین هماهنگی را با دنیای خارج(همان مثال ها) از خود نشان دهد.
- بعبارتی، پروسه یادگیری (Training or Learning) پروسه ای است که در آن یک سیستم پارامتر های آزاد خود را به نحوی تغییر می دهد که بیشترین تطابق را به محیط (نمونه های یادگیری) نشان دهد.

انواع یادگیری

- یادگیری با مربی Supervised Learning
  - مثال ها از زوج های مرتب (ورودی-کلاس) تشکیل شده اند
- یادگیری بدون مربی Unsupervised Learning
  - مثال ها تنها شامل ورودی هستند و توزیع پراکندگی نمونه ها راهنمای عمل خواهد بود
- یادگیری تشدیدی Reinforcement Learning
  - معیاری (معمولا اسکالر) برای رفتار سیستم وجود دارد که راهنمای تغییر در پارامتر هاست.

یادگیری و بهینه سازی

- تطابق با محیط خارجی، توسط معیاری خاص تعریف می شود
  - مثلا بیشترین تصمیم درست روی مجموعه ای از داده ها
- روش بهینه کردن(کمینه کردن و یا بیشینه کردن) آن معیار، تعیین کننده روند یادگیری است.
  - مثلا استفاده از معیار کمترین مربع خطا
- روند مزبور که مبتنی بر جستجو است، می توان یک دفعه ای (Batch) و یا گام به گام(Iterative) باشد.
  - مثلا روش های مبتنی بر گرادیان

تقابل یادگیری و بهینه سازی
تقابل یادگیری و تعمیم

- معلوم نیست بین پروسه یادگیری، که عموما شامل بهینه سازی معیار خاصی است، (مثلا کم کردن معیار مانند مجموع مربعات خطا) و هدف اصلی سیستم (طبقه بندی با کمترین خطای طبقه بندی یا کمترین ریسک) رابطه مستقیمی باشد.
  - علت: توابع هزینه در طبقه بندی گسسته هستند ولی اغلب روش های جستجو در بهینه سازی به توابع پیوسته مربوط شده است
- معلوم نیست بین وصول به هدف سیستم روی نمونه های یادگیری و نمونه هایی که قبلا دیده نشده اند مطابقت وجود داشته باشد (تقابل یادگیری و تعمیم)
  - علت: عدم اطمینان از حل مشکلات یادگیری بخصوص در تطابق با نویز
  - عدم تطابق کامل بین مجموعه یادگیری و تست

حل عملی تقابل یادگیری در برابر تعمیم

- نیاز است که داده ها لااقل به سه بخش تقسیم شود:
  - داده های یادگیری که برای طرح سیستم، مثلا یافتن پارامتر های آزاد مدل، مورد استفاده قرار می گیرند. (training set)
  - (گاهی) بخشی از داده های یادگیری برای تعیین خود پارامتر های فرآیند یادگیری (مثلا زمان توقف آن) نیز مورد استفاده قرار می گیرند که به این بخش از داده های داده های صحت آزمائی می گویند. (validation set)
  - داده های آزمایش که به هیچ وجه در مرحله یادگیری مورد استفاده قرار نگرفته و در واقع آزمایش سیستم طرح شده روی آنها، نشان دهنده کارآیی سیستم طرح شده، روی داده های دیده نشده خواهدبود. (test set) این مسئله به قابلیت تعمیم generalization موسوم است.

انواع کاربرد داده ها (یادگیری) در classification

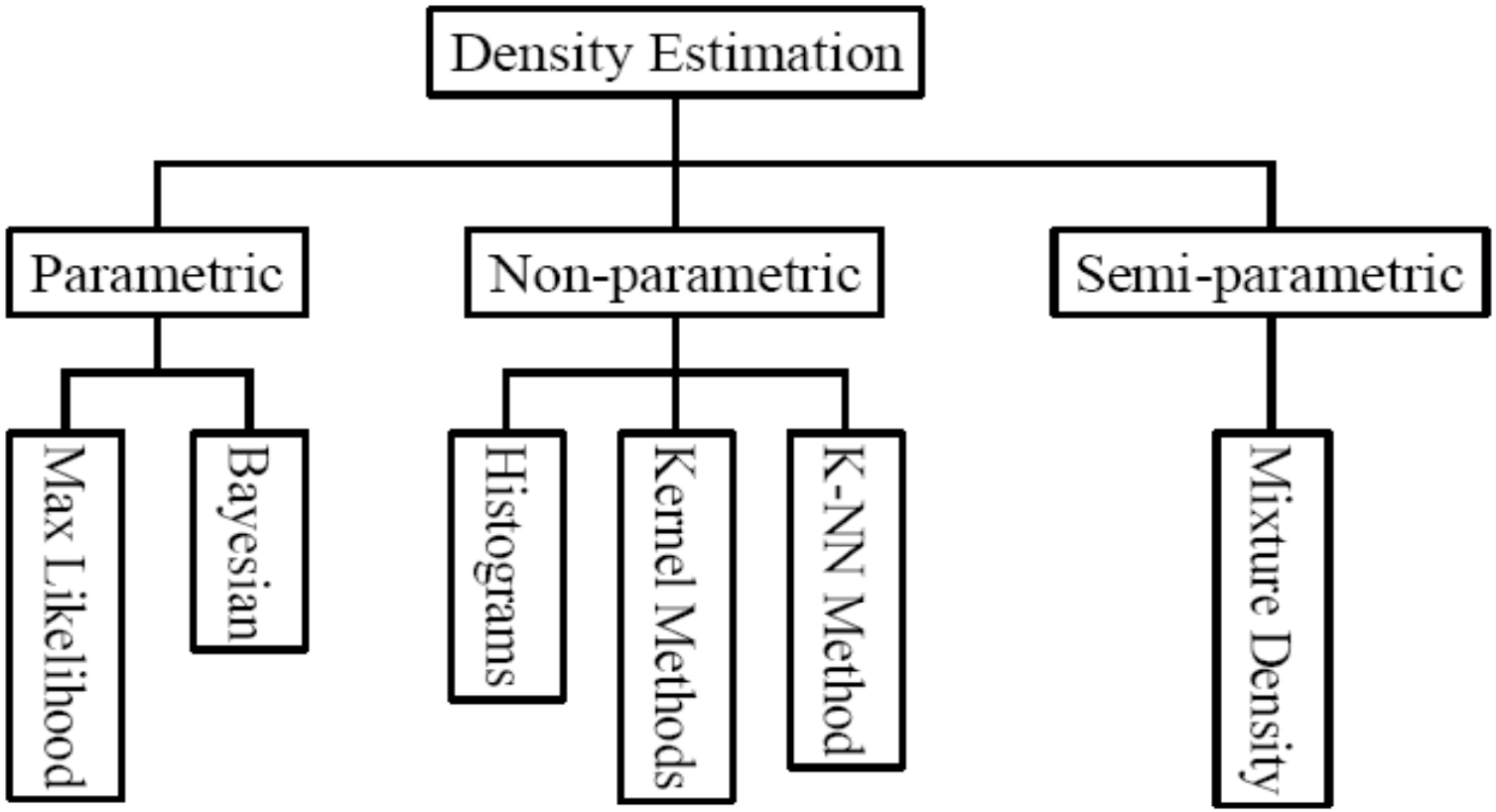
- classification مستقیم
  - طراحی انواع دیگر کلاسیفایر مانند شبکه های عصبی، درخت های تصمیم، بردار های پشتیبان، نزدیکترین همسایگی ....
  - در اغلب این موارد از تشابه یا هم ارزی تلویحی مفاهیم دیگر با احتمال وقوع کلاس استفاده می شود.
- تخمین احتمال پسین وقوع کلاس P(W|x)
  - استفاده برای تعیین عمل (و نه فقط تصمیم) و نیز طبقه بندی سلسله مراتبی و مرکب
- تخمین چگالی احتمال p(x|w)
  - جهت استفاده از ساختار های بهینه مانند کلاسیفایر Bayesian

بحث امروز در این مورد خواهد بود.

تخمین چگالی احتمال با یادگیری با مربی

- در بسیاری موارد ترجیح داده می شود که با تخمین چگالی احتمال از کلاسیفایر Bayesian استفاده شود.
  - برای پیدا کردن توزیع یا احتمال پیشین مشکلی نیست
    - مثلا احتمال پیشین با شمارش تعداد اعضاء هر کلاس در مجموعه داده های یادگیری بدست می آید.
  - اعضای کلاس های مختلف در یادگیری کاملا از هم تفکیک شده اند و ما می توانیم از هر کلاس مستقل از دیگر کلاس ها تجزیه و تحلیل داشته باشیم
  - دو رویکرد اساسی، برای یافتن توابع توزیع احتمال با استفاده از داده ها عبارتند از:
    - روش پارامتری که در آن نوع تابع توزیع معلوم فرض می شود ولی پارامتر هایش نامعلوم است
    - روش غیر پارامتری که در آن شکل تابع نیز نامعلوم است.

تخمین چگالی احتمال



روش های پارامتری تخمین تابع توزیع

- در بسیاری از موارد شکل توابع توزیع پیچیده است اما با دانستن تعداد معدودی اطلاعات کلیدی که به آنها پارامتر می گوئیم می توان تابع را دقیقا مشخص نمود.
- مثلا تابع توزیع گوسین با داشتن مقدار متوسط mean و پراکندگی آن (variance or Covariance matrix) کاملا تعریف می شود.
- بعبارت دیگر کافی است از نمونه ها برای پیدا کردن بردار حاوی این مجهول ها(مثلا با حل دستگاه چند معادله چند مجهولی)، استفاده شود.
- رویکرد ها ی تخمین بردار پارامتر ها:
  - فرض مقدار مطلق (فیکس) ولی نامعلوم برای پارامتر ها (روش بیشترین شباهت یا بیشترین احتمال (Maximum Likelihood)
  - فرض مقداری اتفاقی با توزیع مشخص (و پارامتر های آن توزیع مجهول) (روش Bayesian)

تبیین مسئله-روش بیشترین شباهت

- فرض کنیم که مثال های مربوط به یک کلاس معین در مجموعه D موجود است.

$$D = \{x_1, x_2, \ldots, x_n\}$$

- هدف، پیدا کردن بردار پارامتر های θ به نحوی است که احتمال شباهت زیر را حداکثر نماید:

$$P(x_1,\ldots, x_n \mid \theta) = \Pi^{1,n} P(x_k \mid \theta);\ |D| = n$$

$$\max_{\theta} P(D \mid \theta) = \max P(x_1,\ldots,x_n \mid \theta)$$

$$= \max \prod_{k=1}^{n} P(x_k \mid \theta)$$

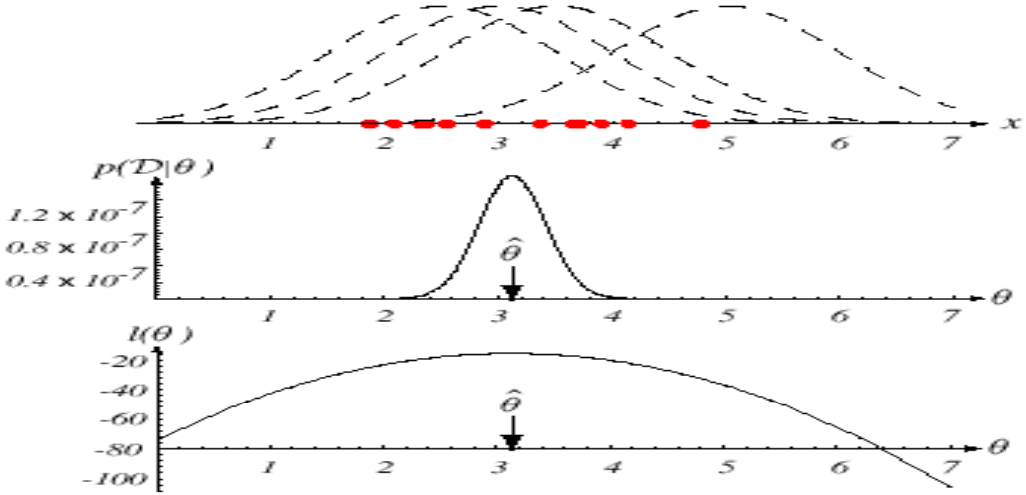


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood—that is, the log-likelihood $l(\theta)$, shown at the bottom. Note that even though they look similar, the likelihood $p(\mathcal{D}|\theta)$ is shown as a function of θ whereas the conditional density $p(x|\theta)$ is shown as a function of x. Furthermore, as a function of θ, the likelihood $p(\mathcal{D}|\theta)$ is not a probability density function and its area has no significance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

تخمین پارامتری-بیشترین شباهت...

- ابتدا برای راحتی تابعی یکنوا از این احتمال تعریف می کنیم(لگاریتم)

$$l(\theta) = \ln P(D \mid \theta)$$

- حال باید مشتق (گرادیان) این تابع را یافته و مساوی صفر قرار دهیم:

$$\hat{\theta} = \arg\max_{\theta} l(\theta)$$

$$(\nabla_{\theta} l = \sum_{k=1}^{k=n} \nabla_{\theta} \ln P(x_k \mid \theta))$$

- مقادیری از بردار پارامتر ها θ که در معادله فوق صدق کند کاندیدائی برای جواب مسئله است

مثال توزیع گوسی با مقدار متوسط مجهول

$$P(x_i \mid \mu) \sim N(\mu, \Sigma)$$

$$\ln P(x_k \mid \mu) = -\frac{1}{2}\ln\left[(2\pi)^d |\Sigma|\right] - \frac{1}{2}(x_k - \mu)^t \Sigma^{-1}(x_k - \mu)$$

$$\text{and} \quad \nabla_{\theta\mu} \ln P(x_k \mid \mu) = \Sigma^{-1}(x_k - \mu)$$

$\theta = \mu$ بنابر این تخمین ML می گوید:

$$\sum_{k=1}^{k=n} \Sigma^{-1}(x_k - \hat{\mu}) = 0$$

ادامه

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{k=n} x_k$$

- با توجه به اینکه ماتریس کوواریانس مخالف صفر است (معکوس دارد) می توان با ضرب آن در دو طرف رابطه به نتیجه ساده تری رسید:
- یعنی تخمین MLاز مقدار mean عبارت است از متوسط حسابی و این مطلب با روش های معمول کاملا همخوانی دارد.

مثال دوم : تخمین متوسط و واریانس برای توزیع گوسی

$$\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$$

$$l = \ln P(x_k \mid \theta) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_\theta l = \begin{pmatrix} \frac{\sigma}{\sigma\theta_1}(\ln P(x_k \mid \theta)) \\ \frac{\sigma}{\sigma\theta_2}(\ln P(x_k \mid \theta)) \end{pmatrix} = 0$$

$$\begin{cases} \frac{1}{\theta_2}(x_k - \theta_1) = 0 \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} = 0 \end{cases}$$

تخمین بیشترین شباهت....

با جمع کردن برای همه نمونه ها:

$$\begin{cases} \sum_{k=1}^{k=n} \frac{1}{\hat{\theta}_2}(x_k - \theta_1) = 0 & (1) \\ -\sum_{k=1}^{k=n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{k=n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 & (2) \end{cases}$$

با ترکیب ۱ و ۲ می توان دریافت:

$$\mu = \sum_{k=1}^{k=n} \frac{x_k}{n} \quad ; \quad \sigma^2 = \frac{\sum_{k=1}^{k=n} (x_k - \mu)^2}{n}$$

ارزیابی تخمین

- در ارزیابی تخمین عموما فاکتورهای مرتبت با دقت در نظر گرفته می شوداما در این راه چند نکته وجود دارد :
  - میزان اختلاف مقادیر تخمین با مقادیر واقعی است.
  - از آنجائیکه عموما مقادیر واقعی در دست نیستند برای محاسبه دقت، با ایجاد تغییراتی در تخمین(داده های تخمین و محاسبات تخمین) میزان تغییرات را می سنجند.
  - میزان صحت پیش فرض ها و حساسیت نسبت به پیش فرض های لازم برای تخمین یکی دیگر از معیار های کارآیی تخمین است
  - اختلاف بین مقادیر تخمین زده شده با مقادیر واقعی می تواند ناشی از دو چیز باشد، مدل(پارامتری) انتخاب شده و داده های انتخاب شده برای تعیین پارامتر ها

مفاهیم بایاس و واریانس

یک روش برای تجزیه و تحلیل خطای تخمین، تجزیه آن به دو بخش است که هر یک به یکی از عوامل یادگیری ارتباط دارد:

- نمونه های بکار گرفته شده، به گونه ای که با تغییر این نمونه ها، پاسخ (و میزان خطا) عوض شود. این بخش از خطا را **واریانس** خطا می گویند.
  - اغلب با افزایش تعداد نمونه ها این بخش از خطا کاهش می یابد.
- مدل و سیستم محاسبه، که ربطی به نمونه ها ندارد و با تغییر آنها عوض نمی شود. این بخش از خطا را **بایاس** می گویند.
- روش ML در تخمین واریانس توزیع گوسی، دارای بایاس است.

بررسی بایاس در تخمین زن ML

- تخمین زن حداکثر شباهت، دارای خطای بایاس زیر است:

$$E\left[\frac{1}{n}\Sigma(x_i - \bar{x})^2\right] = \frac{n-1}{n}.\sigma^2 \neq \sigma^2$$

- یک تخمین زن بدون بایاس برای واریانس بشرح زیر پیشنهاد می شود:

$$\underbrace{C = \frac{1}{n-1}\sum_{k=1}^{k=n}(x_k - \mu)(x_k - \hat{\mu})^t}_{\text{Sample covariance matrix}}$$

تخمین Bayesian

- در تخمین Bayesian فرض می شود که بردار θ (پارامتر های تابع توزیع) نه یک مقدار فیکس مجهول، بلکه حاصل یک فرآیند اتفاقی است.
- توزیع فرآیند بوجود آورنده θ مشخص است اما پارامتر های آن خود مجهول است.
- در اینجا نیز محاسبه احتمال پسین وقوع کلاس ها $P(\omega_i \mid x)$ هسته اصلی محاسبات در پروسه بهینه سازی است.

روش Bayesian در تخمین پارامتری
تبیین مسئله

- آنچه ما خواهیم داشت عبارتند از:
  - نمونه هایی مانند x
  - کلاس هایی مانند ω
  - مجموع داده های یادگیری مانند D
- انتخاب کلاس با دیدن هر نمونه و با در دست داشتن مجموعه داده های یادگیری بر این اساس است که بتوانیم احتمال پسین وقوع کلاس ها را محاسبه نموده و نمونه را به کلاسی نسبت دهیم که دارای بیشترین احتمال پسین باشد.

$$\underset{\omega_j}{Max}\left[P(\omega_j \mid x,\square\right] \equiv \underset{\omega_j}{Max}\left[p(x \mid \omega_j,\square_j).P(\omega_j)\right]$$

احتمال شرطی وقوع کلاس با داشتن مجموعه یادگیری

- میتوان چنین تابعی را به شکل زیر نوشت:

$$P(\omega_i \mid x, \mathsf{D}) = \frac{p(x \mid \omega_i, \mathsf{D}).P(\omega_i \mid \mathsf{D})}{\sum_{j=1}^{c} p(x \mid \omega_j, \mathsf{D}).P(\omega_j \mid \mathsf{D})}$$

- البته باید برای هریک از کمیت های بکار رفته در فرمول بتوان مبنائی برای محاسبه پیدا نمود

روش Bayesian ...

- فرض کنیم که داده های یادگیری را در D جای داده باشیم.
- فرض کنیم همه اطلاعات کلاس در نمونه های آن کلاس قابل دسترسی باشد
- فرض کنیم که توزیع کلاس ها از یکدیگر مستقل قابل محاسبه باشد
- فرض کنیم که نمونه برداری به گونه ای صورت گرفته باشد که احتمال پیشین وقوع کلاس ها از این مجموعه داده ها مستقل بوده و بسادگی می توان نوشت:

$$P(\omega_i) = P(\omega_i \mid \mathrm{D}) \ \text{(Training sample provides this!)}$$

روش Bayesian ...

- حال باید بتوانیم توزیع هر کلاس را پیدا نمائیم.

$$p(x \mid \omega_i, D_i)$$

- به این ترتیب آنچه به دنبال آن هستیم به شرح زیر تبدیل می شود

$$P(\omega_i \mid x, \mathrm{D}) = \frac{p(x \mid \omega_i, \mathrm{D}_i).P(\omega_i)}{\sum_{j=1}^{c} p(x \mid \omega_j, \mathrm{D}).P(\omega_j)}$$

تخمین Bayesian
: Gaussian Case

هدف اولیه : یافتن توزیع $p(\theta \mid D)$

- در حالت یک متغیره به دنبال $p(\mu \mid D)$

μ یک متغیر اتفاقی با توزیع مشخص (مثلا گوسی با پارامتر های مشخص است)

$(\mu_0 \ \& \ \sigma_0)$پارامتر های توزیع متغیر تصادفی θ هستند

$$p(x \mid \mu) \sim N(\mu, \sigma^2)$$

$$p(\mu) \sim N(\mu_0, \sigma_0^2)$$

روال کار

$$P(\omega_i \mid x, D) = \frac{p(x \mid \omega_i, D) P(\omega_i \mid D)}{p(x \mid D)}$$

This is a supervised problem so far:

$$D = \{D_1, D_2, \ldots, D_N\}$$

$$p(x \mid \omega_i, D) = p\left(x \mid \omega_i, \{D_j\}_{j=1\ldots N}\right)$$

$$= p\left(x \mid \omega_i, D_i, \{D_j\}_{j \neq i}\right) = p(x \mid \omega_i, D_i)$$

$$P(\omega_i \mid x, D) = \frac{p(x \mid \omega_i, D_i) P(\omega_i \mid D)}{p(x \mid D)}$$

We will assume that we can obtain "labeled" data, so again:

Notationally: $p(x \mid \omega_i, D_i) \Longrightarrow p(x \mid D)$

Now our problem is to compute density for x given the data D.

We assume the form of $p(x)$ – the source density for D:

$$p(x) \longrightarrow p(x \mid \theta)$$

… and treat θ as a *random variable*

استفاده از تمام توزیع به جای انتخاب تنها یک مقدار

Instead of choosing a value for a parameter, we use them all:

$$p(x \mid D) = \int p(x,\theta \mid D)d\theta = \int p(x \mid \theta, D)p(\theta \mid D)d\theta$$

Data predicts the new sample

x is independent of D given θ

$$= \int \underbrace{p(x \mid \theta)}\underbrace{p(\theta \mid D)}d\theta$$

We chose the form of this

What is this?

Average densities $p(x|\theta)$ for ALL possible values of θ weighted by its posterior probability

ادامه

Computing the posterior probability for θ :

$$\int p(x \mid \theta)\, p(\theta \mid D)\, d\theta$$

Using Bayes rule:

What is this?

Prior belief about the parameters (denisty)

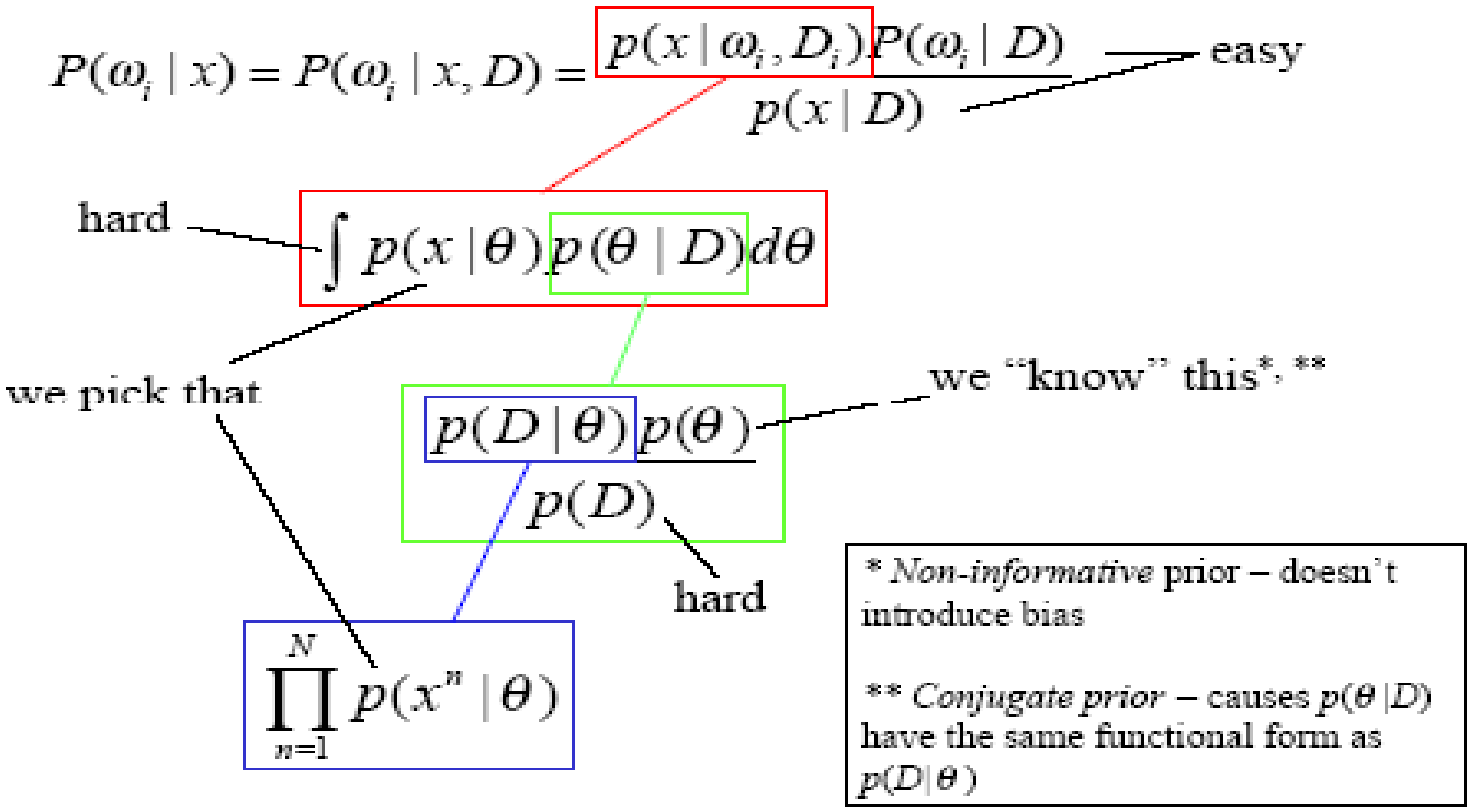
$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{\int p(D \mid \theta)\, p(\theta)\, d\theta}$$

Using independence:

$$p(D \mid \theta) = \prod_{n=1}^{N} p(x^n \mid \theta)$$

Bayesian method does not commit to a particular value of θ, but uses the *entire distribution*.

خلاصه:


$P(\omega_i \mid x) = P(\omega_i \mid x, D) = \frac{p(x \mid \omega_i, D_i) P(\omega_i \mid D)}{p(x \mid D)}$
easy
hard
$\int p(x \mid \theta) p(\theta \mid D) d\theta$
we pick that
$\frac{p(D \mid \theta) p(\theta)}{p(D)}$
we "know" this*, **
hard
$\prod_{n=1}^{N} p(x^n \mid \theta)$
* *Non-informative* prior – doesn't introduce bias
** *Conjugate prior* – causes $p(\theta \mid D)$ have the same functional form as $p(D \mid \theta)$

مثال تخمین Bayesian برای توزیع نرمال

- در حالت یک متغیره :$p(x \mid D)$
  - $P(\mu \mid D)$محاسبه شده
  - $P(x \mid D)$ باید محاسبه شود

$$\mathbf{P(x \mid D) = \int P(x \mid \mu).P(\mu \mid D)d\mu \ \ is\ Gaussian}$$

$$p(x \mid D) \sim N(\mu_n, \sigma^2 + \sigma_n^2)$$

در نتیجه

(Desired class-conditional density $p(x \mid D_j, \omega_j)$)

و قانون Bayes:

$$\underset{\omega_j}{Max}\left[P(\omega_j \mid x, D\right] \equiv \underset{\omega_j}{Max}\left[p(x \mid \omega_j, D_j).P(\omega_j)\right]$$

$$P(\mu \mid \mathbf{D}) = \frac{p(\mathbf{D} \mid \mu).P(\mu)}{\int p(\mathbf{D} \mid \mu).P(\mu)\,d\mu} \qquad (1)$$

$$= \alpha \prod_{k=1}^{k=n} p(x_k \mid \mu).P(\mu)$$

$$p(\mu|\mathbf{D}) \sim N(\mu_n, \sigma_n^2) \qquad (2)$$

$$\mu_n = \left(\frac{n\sigma_0^2}{n\,\sigma_0^2 + \sigma^2} \right) \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}.\mu_0$$

تابع توزیع

$$and \quad \sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}$$

و در نتیجه :

چگونگی روال محاسبه

$$P(\mu|\mathsf{D}) = \frac{p(\mathsf{D}|\mu).P(\mu)}{p(D)} \qquad (1)$$

$$= k_1 \prod_{k=1}^{k=n} p(x_k|\mu).P(\mu) \qquad \mathrm{k1} = \frac{1}{\mathrm{p(D)}}$$

subsituating $P(\mu|\mathsf{D})$ which have normal distribution :

$$= k_2 \exp[\frac{-1}{2}\{[\sum_{k=1}^{n} (\frac{\mu - x}{\sigma})^2] + (\frac{\mu - \mu_0}{\sigma_0})^2\}]$$

$$= k_3 \exp\{ \frac{-1}{2}(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2})\mu^2 - 2[\frac{\sum_{k=1}^{n} x_k}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}]\mu\}$$

ادامه

- حال اگر آنچه بدست آمده است را با شکل کلی یک توزیع نرمال مساوی قرار دهیم می توانیم پارامتر ها را بطور متناظر بدست آوریم

$$p(\mu) = p(\mu \mid D_i) = \frac{1}{\sqrt{2\pi\sigma_n}} \exp\{\frac{-1}{2}(\frac{\mu - \mu_n}{\sigma_n})^2\}$$

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

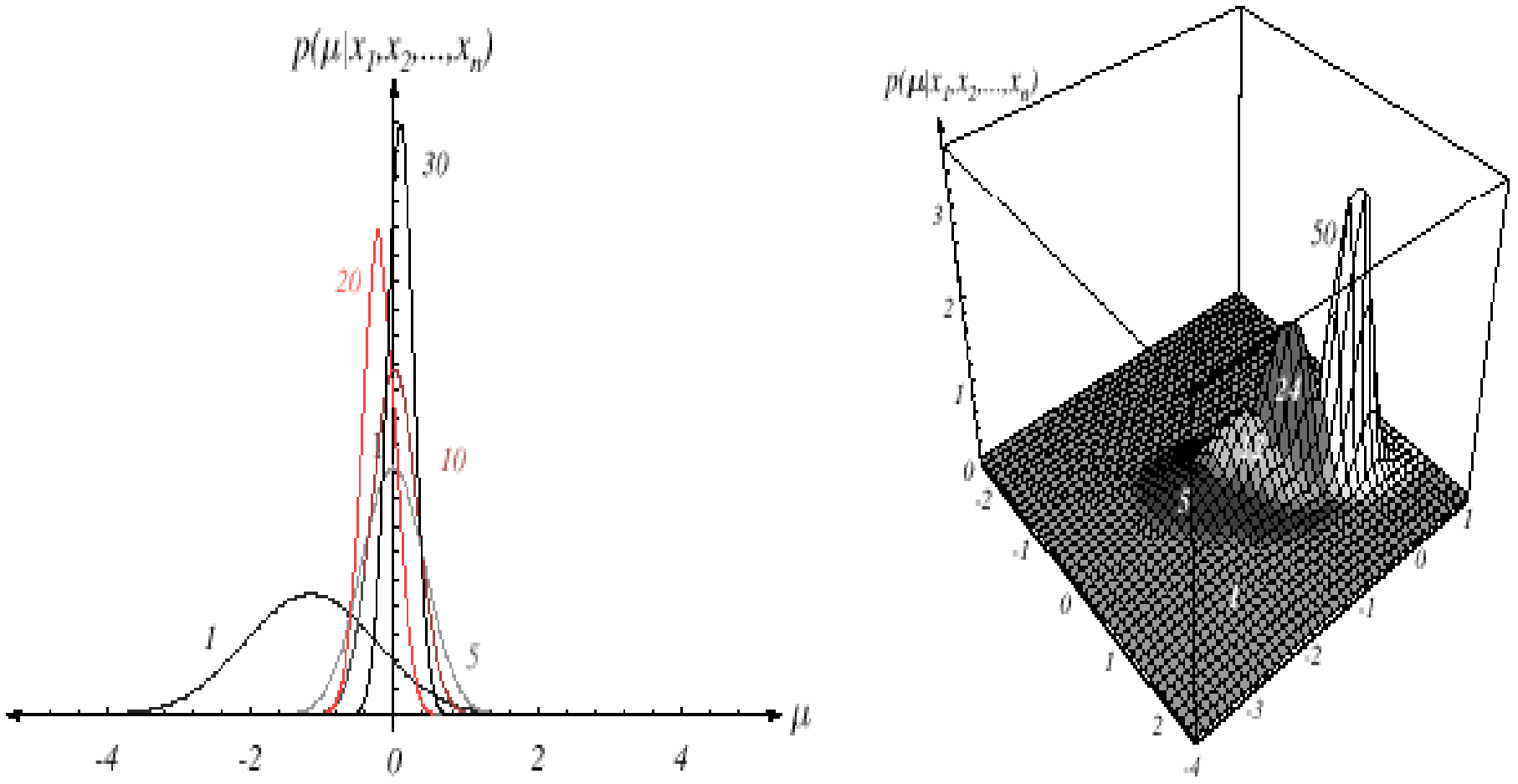


FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

مقایسه و ارتباط بین دو روش

$$p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta)$$

$$= \left[\prod_n p(x^n \mid \theta) \right] p(\theta) = L(\theta)\, p(\theta)$$

peaks at $\hat{\theta}_{ML}$

If the peak is sharp and $p(\theta)$ is flat, then:

$$p(x \mid D) = \int p(x \mid \theta)\, p(\theta \mid D) d\theta$$

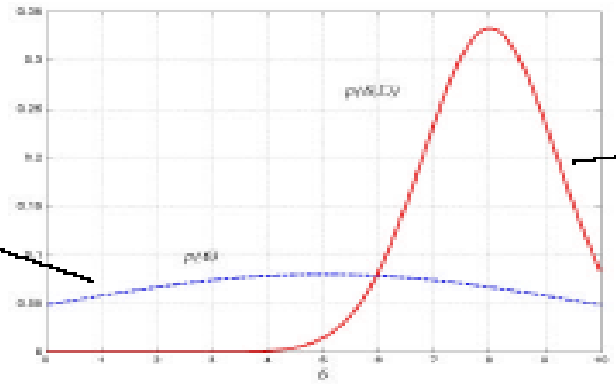
$$\simeq \int p(x \mid \hat{\theta})\, p(\theta \mid D) d\theta = p(x \mid \hat{\theta}) \int p(\theta \mid D) d\theta = p(x \mid \hat{\theta})$$

$$\text{As } N \to \infty,\ p(x \mid D) \leftrightarrow p(x \mid \hat{\theta})$$

ادامه

For $\theta = \mu$:

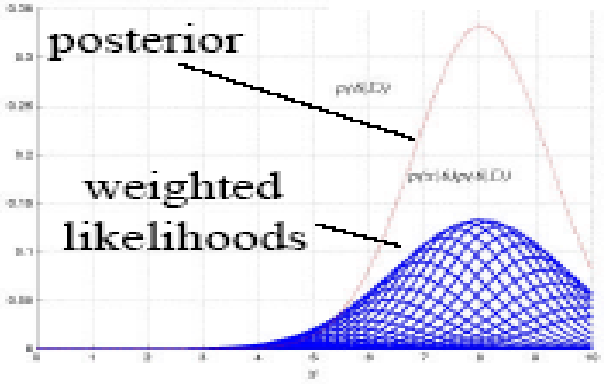
$$\int p(x \mid \mu) p(\mu \mid D) d\mu$$

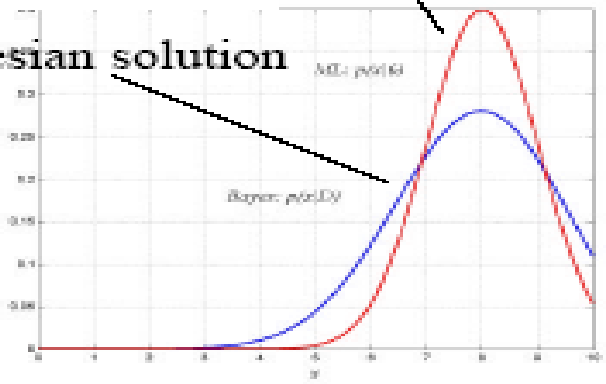


Parameter posterior

Parameter prior

ML solution





اشکال روش Bayesian

- دشواری در پیاده سازی
- در برخی(بیشتر) موارد یافتن جواب ممکن نیست

خلاصه بحث

- سه کاربرد اساسی مثال ها
  - یافتن مستقیم labelها
  - یافتن احتمال پسین عضویت در کلاس ها
  - یافتن توزیع(چگالی) اعضای هر کلاس
- سه رویکرد اساسی در یافتن توزیع از روی نمونه ها
  - یادگیری با مربی
  - یادگیری بدون مربی
  - یادگیری تشدیدی
- دو پیش فرض اساسی:
  - تخمین پارامتری
  - تخمین غیر پارامتری

ادامه خلاصه

- معمولا تخمین پارامتری تابع توزیع یه یک فرایند بهینه سازی منجر می شود که در آن یک تابع هدف یا هزینه با یستی به بیشترین یا کمترین مقدار خود برسد.
- دو روش معمول در تخمین پارامتری
  - با فرض اینکه پارامتر ها تعیینی ولی مجهولند روش Maximum Likelihood (ML)
  - با فرض اینکه پارامتر ها خروجی پروسه های اتفاقی هستند روش Bayesian

پایان

سوال؟