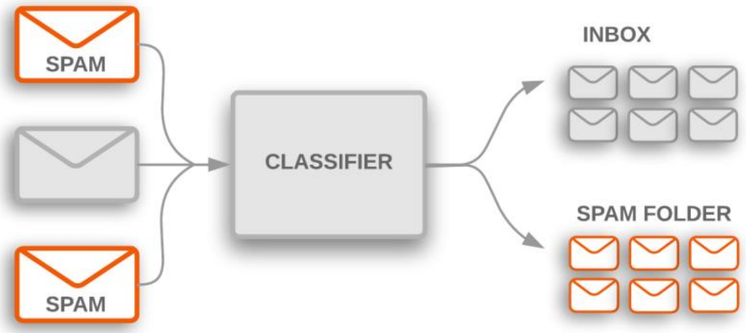
طبقه‌بندی
(Classification)

طبقه‌بندی کردن داده‌ها

- تعریف
  - در مسئله رگرسیون خطی، بردار خروجی یک بردار حقیقی است.
  - در طبقه بندی کردن داده ها بردار خروجی یک بردار گسسته است.



- هدف
  - اختصاص دادن یک مقدار گسسته به یک داده ورودی ← Discriminant function
  - اختصاص دادن احتمال رخ داد یک نمونه در یک کلاس معین ← Probabilistic modeling

طبقه‌بندی کردن داده‌ها

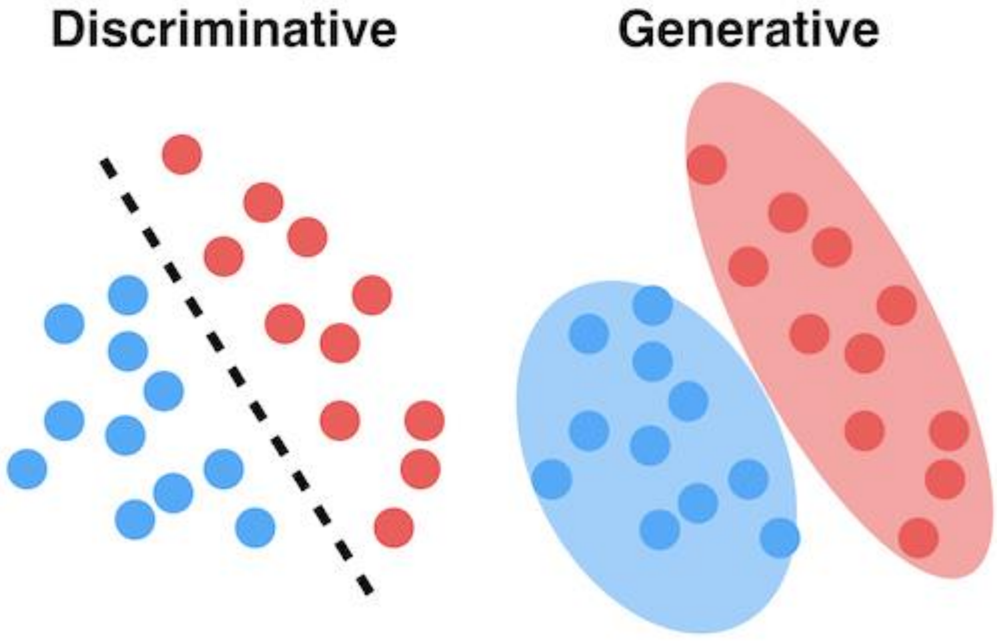
- تشخیص anomaly

y	x	class
1	0	normal
2	0	abnormal
3	1	normal
1	1	normal
0.5	3	normal
2	2.5	abnormal
3	2.4	normal
1	2.2	abnormal
0.4	0.7	abnormal
0.8	0	abnormal

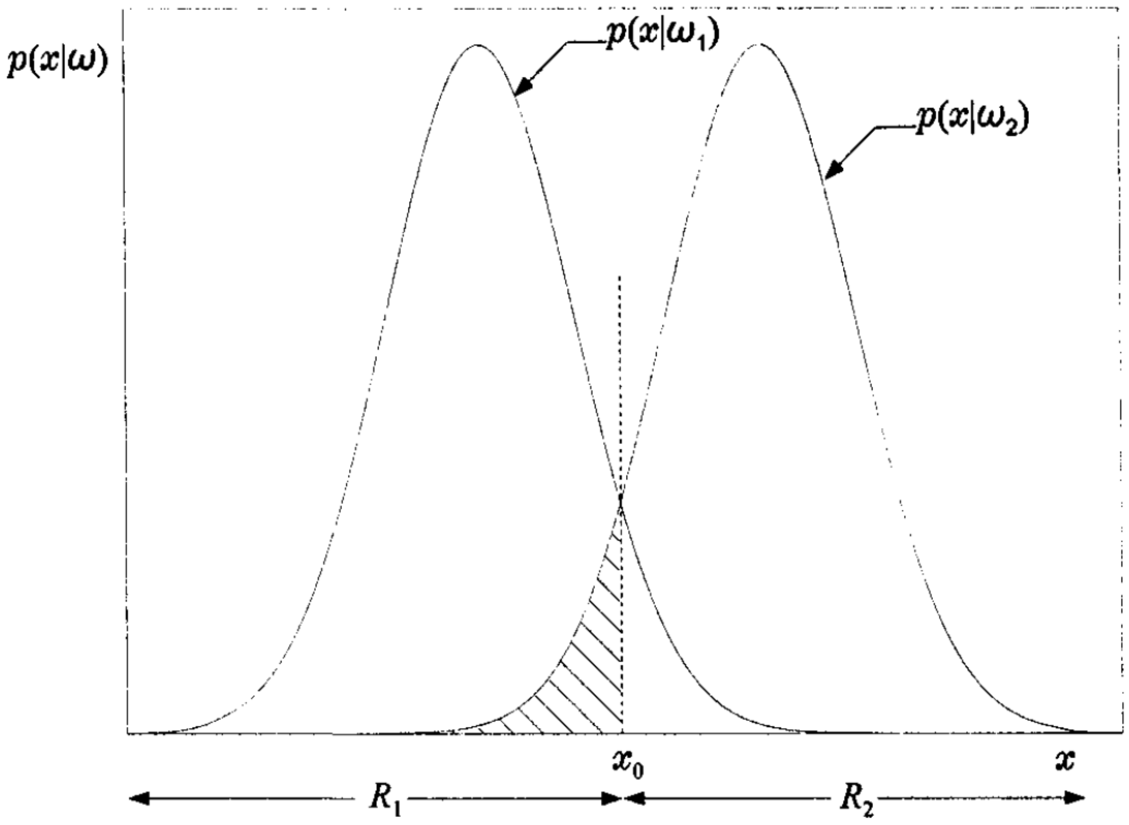
$X = \begin{pmatrix}1.5\\0.5\end{pmatrix}$ ⇨ **Class = ?**

مدل‌های احتمالاتی

- مدل‌های احتمالاتی – تصمیم‌گیری برمبنای قاعده تصمیم‌گیری بیز شکل می‌گیرد.
  - مدل‌های مولد
  - مدل‌های جداکننده



قاعده تصمیم‌گیری بیز



- قاعده تصمیم‌گیری بیز

$$P_e = \int P(w_1)P(x|w_1)dx_{R_2} + \int P(w_2)P(x|w_2)dx_{R_1}$$

$$\int P(w_1)P(x|w_1)dx_{R_1} + \int P(w_1)P(x|w_1)dx_{R_2} = 1$$

$$P_e = 1 - \int (P(w_1)P(x|w_1) - P(w_2)P(x|w_2)dx_{R_1}$$

مدل‌های احتمالاتی مولد

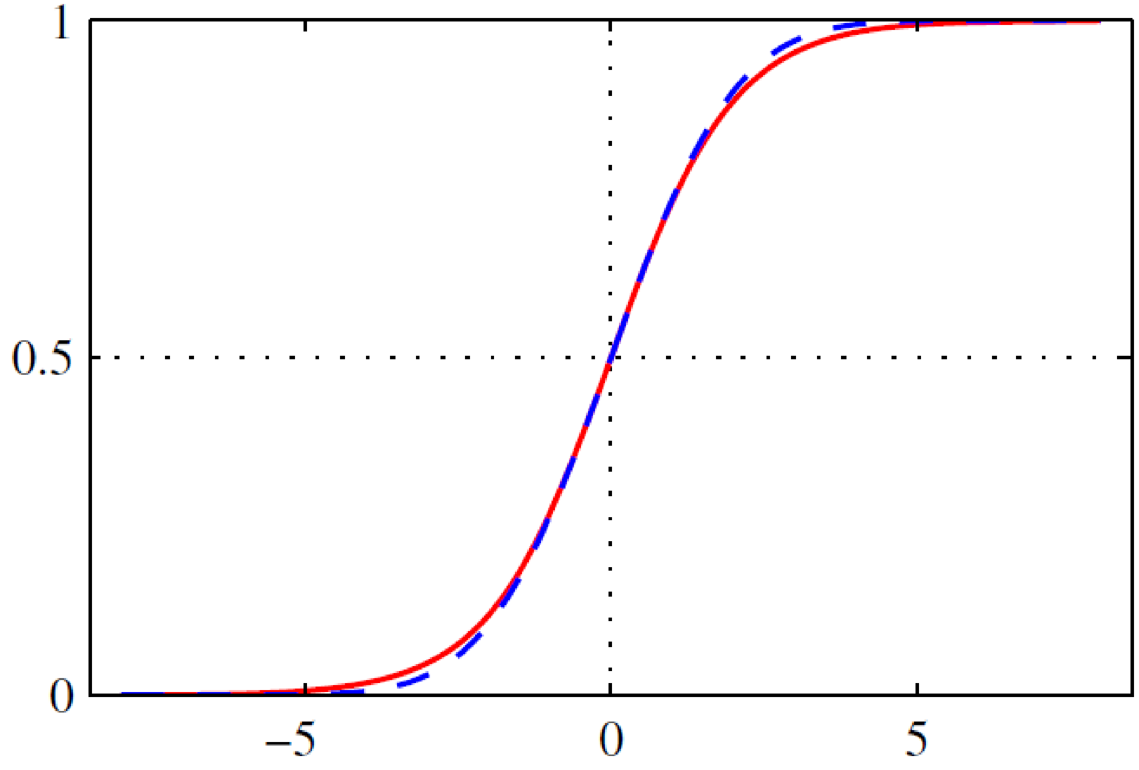
- مدل‌های احتمالاتی مولد
  - توزیع داده‌های هر کلاس و توزیع پیشین هر کلاس اهمیت دارد.

$$P(C_k|x)=\frac{P(x|C_k)P(C_k)}{\sum_{C_k}P(x|C_k)P(C_k)} \Rightarrow P(C_1|x)=\frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)+P(x|C_1)P(C_1)}$$

$$P(C_1|x)=\frac{1}{\frac{P(x|C_2)P(C_2)}{P(x|C_1)P(C_1)}+1}=\frac{1}{1+\exp(-a)}\quad ,a=\ln\frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)}$$

مدل‌های احتمالاتی مولد

- تابع سیگموئید



$$P(C_1|x) = \frac{1}{1+\exp(-a)}, a = \ln\frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)}$$

مدل‌های احتمالاتی مولد

- مدل‌های احتمالاتی مولد

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{\sum_{C_k} P(x|C_k)P(C_k)} \Rightarrow P(C_k|x) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

$$a_k = \ln P(x|C_k)P(C_k)$$

مدل‌های احتمالاتی مولد

- کلاس پیوسته

$$P(x|C_k) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_\mathrm{k}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_\mathrm{k}^{-1}(x-\mu_k)\right)$$

$$a = \ln\frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)} = -\frac{1}{2}\{(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - (x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\}$$

$$a(x) = w^T x + w_0 \Rightarrow \begin{cases} w = \Sigma^{-1}(\mu_1 - \mu_2) \\ w_0 = -\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2\Sigma^{-1}\mu_2 + \ln\frac{P(C_1)}{P(C_2)} \end{cases}$$

مدل‌های احتمالاتی مولد

- پارامترهای مدل هر کلاس به چه صورت انتخاب می‌شوند؟
  - بیشینه شباهت
  - خروجی برای یک مسئله دو کلاسه از چه توزیعی تبعیت می‌کند؟

$$P(x, C_1) = P(C_1)P(x|C_1) = \pi P(x|C_1), \qquad P(x, C_2) = P(C_2)P(x|C_2) = (1-\pi)P(x|C_2)$$

$$J = -\ln P(t|x) = -\ln \prod_{i=1}^{N} \left(\pi P(x_i|C_1)\right)^{t_i} \left((1-\pi)P(x|C_2)\right)^{1-t_i}$$

مدل‌های احتمالاتی مولد

- پارامترهای مدل

$$J = -\ln P(t|x) = -\ln \prod_{i=1}^{N} \left(\pi P(x_i|C_1)\right)^{t_i} \left((1-\pi)P(x|C_2)\ \right)^{1-t_i}$$

$$\pi = \frac{N_1}{N_1 + N_2}$$

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} x_i, \qquad \hat{\Sigma}_k = \frac{1}{N} \sum_{i=1}^{N_k} (x - \hat{\mu}_k)(x - \hat{\mu}_k)^T$$

مقدمه‌ای بر تئوری تخمین

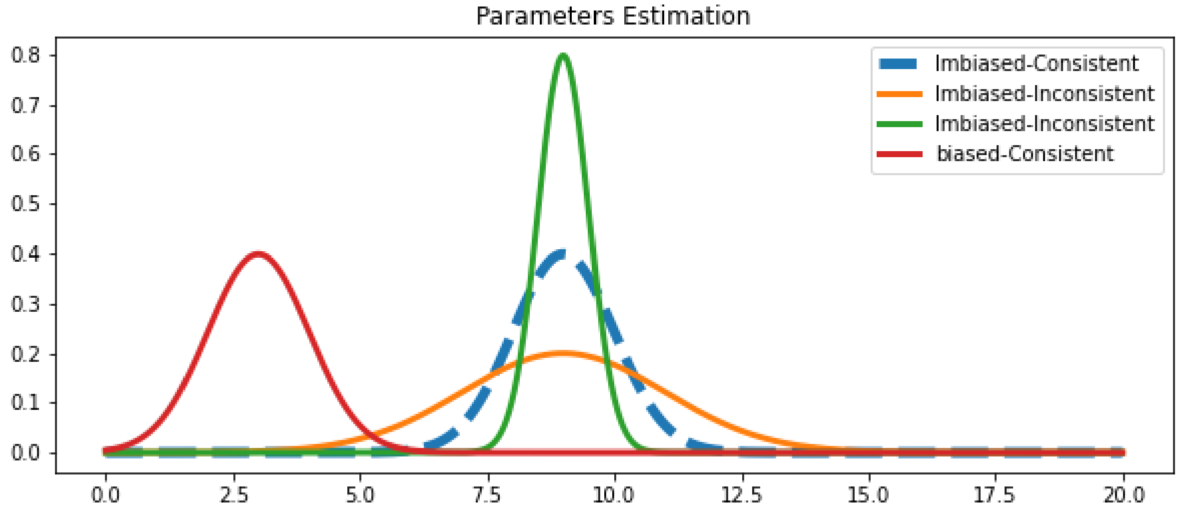
تئوری تخمین

- از آنجا که تخمین پارامترهای توزیع، بر اساس N متغیر تصادفی است، خود نیز متغیر تصادفی است.

$$\hat{\mu}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} x_i, \qquad \hat{\Sigma}_k = \frac{1}{N}\sum_{i=1}^{N_k}(x-\hat{\mu}_k)(x-\hat{\mu}_k)^T$$

$$E(\hat{\mu}_k) = \frac{1}{N_k}\sum_{i=1}^{N_k} E(x_i) \rightarrow E(\hat{\mu}_k) = \mu_k\ ,$$

$$\Sigma_{\hat{\mu}} = \frac{1}{N}\Sigma$$



مدل‌های احتمالاتی مولد

- پارامترهای مدل
    - مدل گوسی – تعداد پارامترهایی که باید تخمین زده شود، متناسب با ابعاد داده‌ها به صورت نمایی تغییر می‌کند.
- Naïve Bayes
    - به شرط مشخص بودن برچسب داده، ابعاد از یکدیگر مستقل هستند.
    - برای مدل گوسی

$$\Sigma = \sigma^2 I$$

مدل‌های احتمالاتی مولد

- کلاس گسسته
  - فرض می‌کنیم داده‌های هر کلاس از یک توزیع برنولی تبعیت می‌کند.
  - اگر بعد ورودی D باشد، برای محاسبه <u>**توزیع جرم احتمال**</u> چند حالت مستقل باید بررسی شود؟

$$P(x|C_k) = \prod_{i=1}^{N} \mu_{ki}^{x_i}(1-\mu_{ki})^{1-x_i}$$

$$a_k(x) = \ln P(C_k)P(x|C_k) = \ln P(C_k) + \sum_{i=1}^{N}\{x_i \ln \mu_{ki} + (1-x_i)\ln(1-\mu_{ki})\}$$

مدل‌های احتمالاتی مولد

- خانواده نمایی Exponential family
  - فرم کلی توزیع‌های این خانواده به صورت زیر است:

$$f(x|\lambda_k) = h(x)g(\lambda)\exp\left(\lambda_k^T u(x)\right)$$

$$a(x) = (\lambda_1 - \lambda_2)^T x + \ln g(\lambda_1) - \ln g(\lambda_2) + \ln P(C_1) - \ln P(C_2)$$

$$a_k(x) = \lambda_k^T x + \ln g(\lambda_k) + \ln P(C_k)$$

مدل‌های احتمالاتی مولد

- بیشینه احتمال موخر
  - آیا تخمین پارامتر صرفا بر اساس مشاهدات رویکرد مناسبی است؟
  - برای یک مدل گوسی

$$\mu^* = \arg\min P(\mu)P(x|\mu) \quad \Longrightarrow \quad P(\mu) = ?$$

  - Conjugate prior
  - برای یک توزیع گوسی، توزیع پیشین گوسی برای میانگین یک مزدوج پیشین است.
  - برای یک توزیع گوسی، توزیع پیشین گاما برای واریانس یک مزدوج پیشین است.

مدل‌های احتمالاتی مولد

- جمع‌بندی

مدل‌های احتمالاتی جداکننده

مدل‌های احتمالاتی جداکننده

- رگرسیون لوجستیک

$$P(C_1|\phi(x)) = y\big(\phi(x)\big) = \sigma(a), \qquad a = W^T\phi(x)$$

$$J = -\ln P(y|\phi(x), W) = -\ln \prod_{i=1}^{N} y_i^{t_i}(1-y_i)^{1-t_i} \Rightarrow$$

$$J = -\sum_{i=1}^{N}\{t_i \ln y_i + (1-t_i)\ln(1-y_i)$$

Cross entropy

- تعداد پارامترهای قابل تنظیم مسئله

آنالیز ریسک

- آنالیز ریسک
  - خطای دسته‌بندی برای هر دو مقدار خطا اهمیت یکسانی قائل می‌شود.
  - آیا خطای دسته‌بندی همواره معیار مناسبی است؟

$$r_k = \sum_i \lambda_{ki} \int f(x|C_k) dx_{R_i} \rightarrow r = \sum_k r_k P(C_k)$$

$$l_i = \sum_k \lambda_{ki} P(x|C_k) P(C_k) < l_j = \sum_k \lambda_{kj} P(x|C_k) P(C_k)$$

آنالیز ریسک

- آنالیز ریسک
  - مسئله دو کلاسه

$$r_1 = \lambda_{11}P(x|C_1)P(C_1) + \lambda_{21}P(x|C_2)P(C_2)$$
$$r_2 = \lambda_{12}P(x|C_1)P(C_1) + \lambda_{22}P(x|C_2)P(C_2)$$

$$r_2 > r_1 \rightarrow (\lambda_{21} - \lambda_{22})P(x|C_2)P(C_2) < (\lambda_{12} - \lambda_{11})P(x|C_1)P(C_1)$$

- کاربرد – تشخیص بیماری

روش‌های غیرپارامتریک

روش‌های غیرپارامتریک

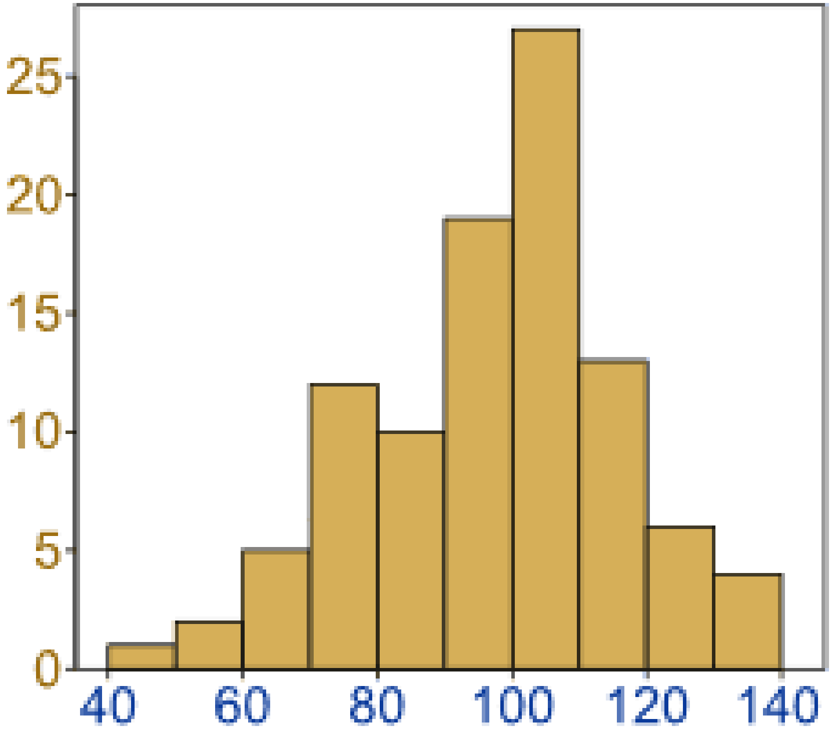
- روش‌های مبتنی بر حافظه
  - آیا داده‌های آموزش پس از آموزش مدل به کار می‌آیند؟

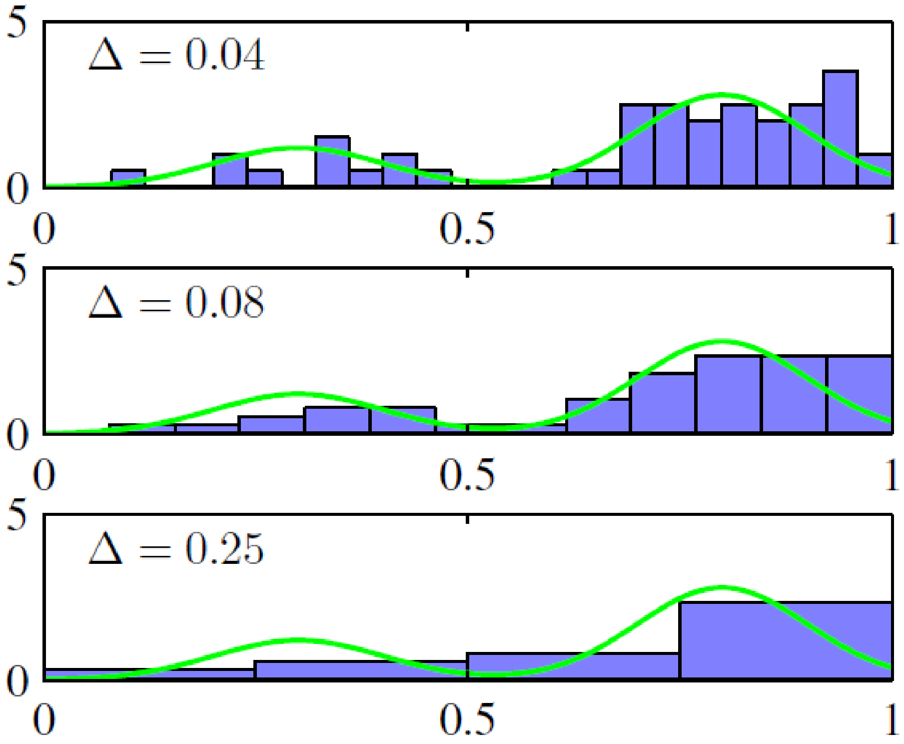
- کرنل
  - تابعی برای سنجیدن فاصله، یا شباهت، بین دو متغیر

روش‌های غیرپارامتریک

- هیستوگرام

$$P = \frac{n_i}{N\Delta_i}$$





روش‌های غیرپارامتریک

- هیستوگرام
- قابلیت تعمیم‌دهی به ابعاد بالا را ندارد.
- به صورت محلی در مورد نمونه در دست تصمیم‌گیری می‌کند.
- نیازمند معیاری برای متر کردن محلی بودن
- تخمین مبتنی بر کرنل
- کرنل پارزن
- نزدیک‌ترین همسایگی

تخمین مبتنی بر کرنل

- احتمال رخداد k نمونه از N نمونه مشاهده شده در ناحیه R چقدر است؟

$$P = \frac{K}{N}$$

$$P = \int p(x)dx$$

$$\Rightarrow$$

$$p(x) = \frac{K}{NV}$$

- تعداد مشاهدات k ثابت باشد.
- حجم ناحیه در نظر گرفته شده ثابت باشد.

روش پارزن

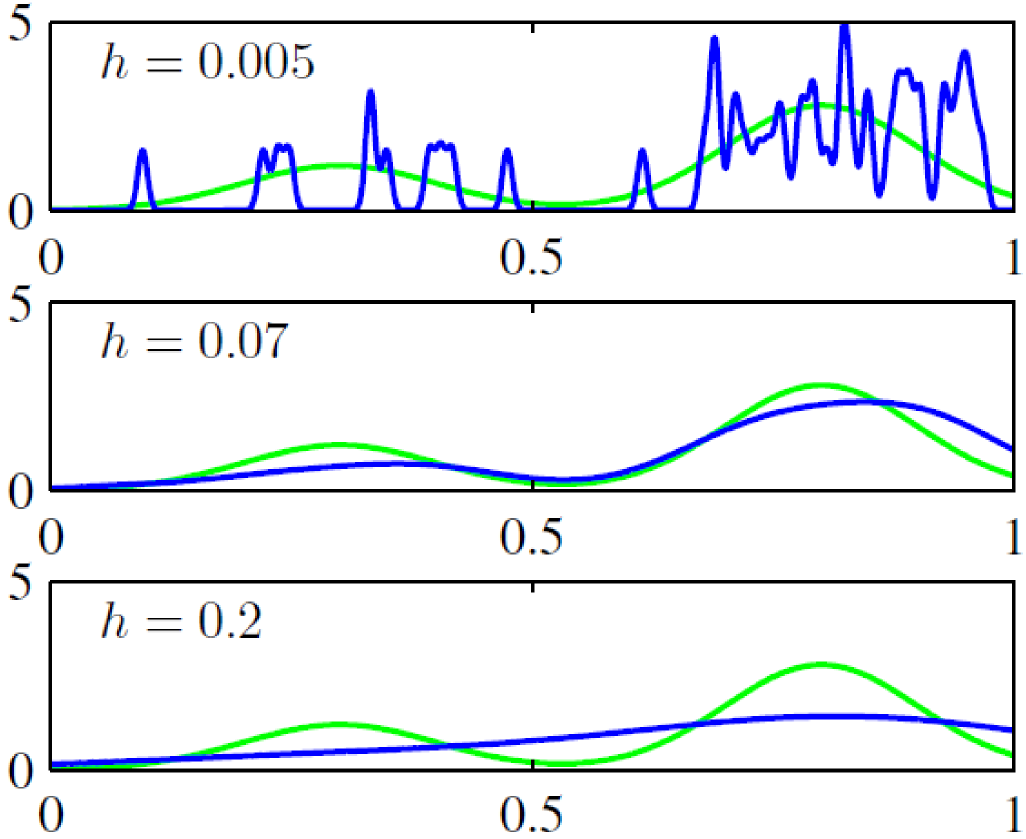
- حجم ناحیه مورد مطالعه ثابت باشد.

$$k(u_i) = \begin{cases} 1, & |u_i| \le \frac{1}{2} \\ 0, & O.W \end{cases} \Rightarrow p(x) = \frac{1}{N}\sum_n \frac{1}{h^D} k\left(\frac{x - x_n}{h}\right)$$

$$p(x) = \frac{1}{N}\sum_n \frac{1}{(2\pi h^2)^{\frac{1}{2}}} \exp\left(-\frac{|x - x_n|^2}{2h^2}\right)$$

- آیا می‌توان هر کرنلی استفاده کرد؟

روش پارزن



- پارامتر h چگونه انتخاب شود؟
  - ماتریس اطمینان

نزدیک‌ترین همسایگی

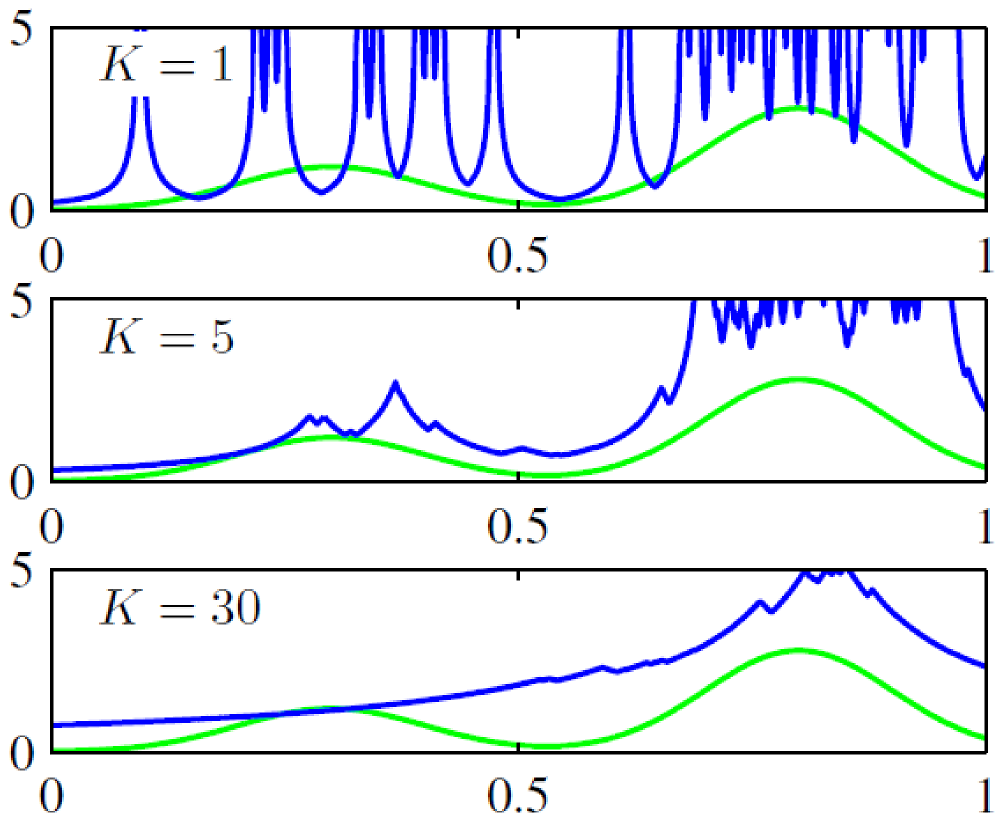
- تعداد مشاهدات k ثابت باشد.
  - ایراد روش کرنل پارزن در چیست؟
- آیا روش نزدیک‌ترین همسایگی تابع چگالی احتمال واقعی را بدست می‌دهد؟

$$p(x) = \frac{K}{NV} \Longrightarrow \int_{-\infty}^{+\infty} \frac{K}{Ndx} = \infty$$

$$P(x|C_k) = \frac{K_k}{N_k V}, \qquad P(C_k) = \frac{N_k}{N}, \qquad P(x) = \frac{K}{NV} \rightarrow P(C_k|x) = \frac{K_k}{K}$$

نزدیک‌ترین همسایگی

- مقدار بهینه k چگونه انتخاب می‌شود؟
  - ماتریس اطمینان



روش‌های غیرپارامتریک

- جمع‌بندی