Linear Discriminant

جداسازهای خطی

یادگیری ماشین

دانشگاه شهید بهشتی
پژوهشکده‌ی فضای مجازی
پاییز ۱۴۰۱
احمد محمودی ازناوه

فهرست مطالب

- انواع دسته‌بندی
  - بر پایه‌ی درست‌نمایی
  - بر پایه‌ی جداساز
- جداساز خطی
  - تعمیم مدل‌های خطی
  - تعبیر هندسی جداساز خطی
  - جداسازی دودسته‌ای و چنددسته‌ای
- مروری بر جداسازی پارامتری
- نزول گرادیان
- Logestic Regression
- Softmax Regression(Multinomil Logestic Regression)
- رتبه‌بندی

دسته‌بندی بر پایه‌ی تابع درست‌نمایی

Likelihood-based classification (Generative models)

- برای دسته‌بندی یک سری «تابع جداساز» ($g_i(x)$) محاسبه می‌شود:

$$g_i(\boldsymbol{x}) = \max_{j=1}^{K} g_j(\boldsymbol{x})$$

اگر

کلاس C_i انتخاب می‌شود

- ابتدا احتمال پیشین($p(C_i)$) و تابع چگالی احتمال داده‌ها در هر کلاس($p(\boldsymbol{x}|C_i)$) بر اساس تابع درست‌نمایی محاسبه شده، سپس بر اساس قانون Bayes، احتمال پسین($P(C_i|\boldsymbol{x})$) به دست آمده و براساس آن تابع جداساز تعریف می‌شود:

$$g_i(\boldsymbol{x}) = \log P(C_i|\boldsymbol{x})$$

- این شیوه به «دسته‌بندی بر پایه‌ی درست‌نمایی» موسوم است. در واقع بر اساس مدلی که برای داده‌ها تخمین زده می‌شود، جداساز به دست می‌آید.
  - در روش‌های پارامتری، نیمه‌پارامتری و ناپارامتری از این شیوه استفاده می‌شود.

دسته‌بندی بر پایه‌ی جداساز

discriminant-based classification (discriminative models)

- در این روش، بدون تخمین توزیع داده‌ها، به صورت مستقیم جداساز تخمین زده می‌شود.
  - در این حالت یک مدل برای جداساز تعریف می‌شود:

$$g_i(\boldsymbol{x}|\Phi_i)$$

  - پارامترهای جداساز به صورت «صریح» مشخص شده‌اند برای خلاف روش‌های مبتنی بر درست‌نمایی که به صورت «ضمنی» و براساس توزیع داده‌ها به دست می‌آیند.

دسته‌بندی بر پایه‌ی جداساز(ادامه...)

- فرآیند آموزش یافتن(بهینه‌سازی) پارامترهای جداساز بر اساس یک مجموعه‌ی آموزشی و با هدف افزایش درستی دسته‌بندی است. در این حالت به جای تخمین درست توزیع داده‌ها هر کلاس، هدف تخمین درست مرزهای بین دسته‌هاست.
- از نظر طرفداران این رویکرد برآورد توزیع داده‌های یک کلاس از تخمین جداساز، مساله‌ی دشوارتری است.
  - برای یک حل مسأله، معقول نیست آن را به مسائل دشوارتر تقسیم کرد!
  - البته زمانی این گفته درست است که جداساز با یک مدل ساده تخمین زده شود.

جداساز خطی

- ساده‌ترین جداسازی که می‌توان در نظر گرفت، «جداساز خطی» است:

$$g_i(\mathbf{x}|\mathbf{w}_i,w_{i0})=\mathbf{w}_i^T\mathbf{x}+w_{i0}=\sum_{j=1}^{d}w_{ij}x_j+w_{i0}$$

- در بیشتر موارد استفاده از جداساز خطی ترجیح داده می‌شود، به دلایل زیر:
  - سادگی: دارای پیچیدگی از مرتبه‌ی $o(d)$
  - تفسیرپذیری: به راحتی می‌توان بر اساس آن به استخراج دانش پرداخت؛ خروجی مجموع وزن‌دار ورودی است. وزن هر بعد اهمیت و علامت آن اثر آن خصیصه را نشان می‌دهد.
  - در بسیاری موارد جداساز خطی بهینه است: توزیع داده‌های کلاس گاوسی با ماتریس کواریانس یکسان

تعمیم مدل‌های خطی

- چنان‌که مدل خطی پیچیدگی لازم را نداشته باشد، می‌توان سراغ جداسازهای پیچیده‌تری رفت:

$$g_i(\mathbf{x}|\mathbf{W}_i,\mathbf{w}_i,w_{i0})=\mathbf{x}^T\mathbf{W}_i\mathbf{x}+\mathbf{w}_i^T\mathbf{x}+w_{i0}$$

جداساز درجه‌ی دو

$$g_i(\mathbf{x})=w_0+\sum_{i=1}^{k}w_i x_i+\sum_{i=1}^{k}\sum_{j=1}^{k}w_{ij}x_i x_j$$

  – دارای پیچیدگی از مرتبه‌ی $o(d^2)$
  – نیاز به داده‌های آموزشی بیشتر
  – احتمال بروز overfitting بیشتر

- یک راه معادل استفاده از جملات با مرتبه بالاتر (higher order terms) است.

$$z_1=x_1,\; z_2=x_2,\; z_3=x_1^2,\; z_4=x_2^2,\; z_5=x_1x_2$$

  – به جای جداساز پیچیده نگاشت غیر خطی به فضایی با جداساز خطی

$$g_i(\mathbf{x})=\sum_{j=0}^{k}w_{ij}\boxed{\varphi_j}(\mathbf{x})$$

Potential functions(1964)

Basis functions

Two Classes

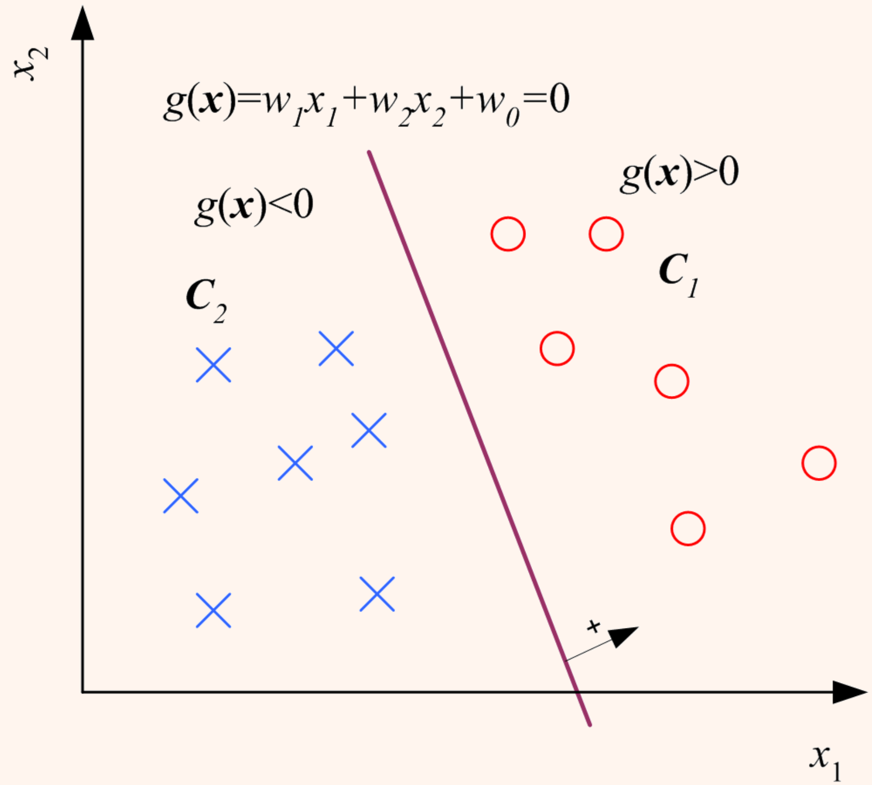
دسته‌بندی دوتایی

- در این حالت یک تابع جداساز ([ابر]سطح جداکننده) کافیست:

$$
\begin{aligned}
g(\mathbf{x}) &= g_1(\mathbf{x}) - g_2(\mathbf{x}) \\
&= \left(\mathbf{w}_1^T\mathbf{x} + w_{10}\right) - \left(\mathbf{w}_2^T\mathbf{x} + w_{20}\right) \\
&= \left(\mathbf{w}_1 - \mathbf{w}_2\right)^T\mathbf{x} + \left(w_{10} - w_{20}\right) \\
&= \boxed{\mathbf{w}^T}\mathbf{x} + \boxed{w_0}
\end{aligned}
$$

بردار وزن‌ها

حدآستانه



$$
\text{choose} \begin{cases} C_1 & \text{if } g(\mathbf{x}) > 0 \\ C_2 & \text{otherwise} \end{cases}
$$

دسته‌بندی دوتایی-تعبیر هندسی

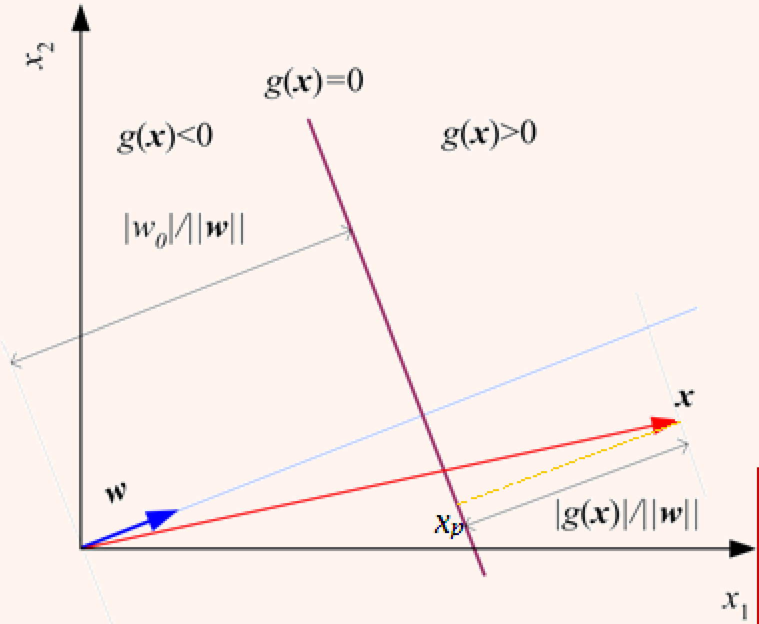
- در صورتی که r فاصله‌ی x از [ابر]سطح جداکننده باشد و x_p نگاشت x بر روی سطح، خواهیم داشت:

$$\boldsymbol{x} = \boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$$

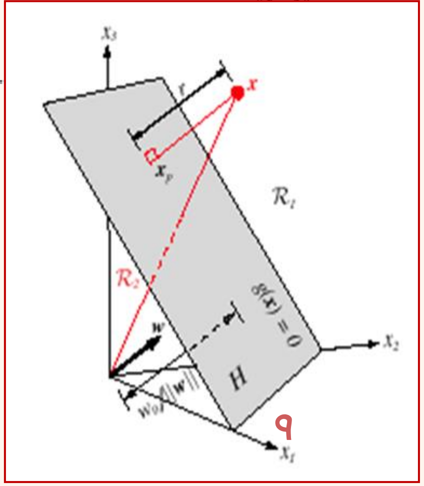
$$g(\boldsymbol{x}) = \boldsymbol{w}^T\left[\boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right] + w_0$$

$$g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}_p + w_0 + r\frac{\boldsymbol{w}^T\boldsymbol{w}}{\|\boldsymbol{w}\|}$$

$$g(\boldsymbol{x}) = r\|\boldsymbol{w}\| \qquad r = \frac{g(\boldsymbol{x})}{\|\boldsymbol{w}\|} \qquad r_0 = \frac{w_0}{\|\boldsymbol{w}\|}$$



Duda et al.



دسته‌بندی چندگروهی

one-versus-all

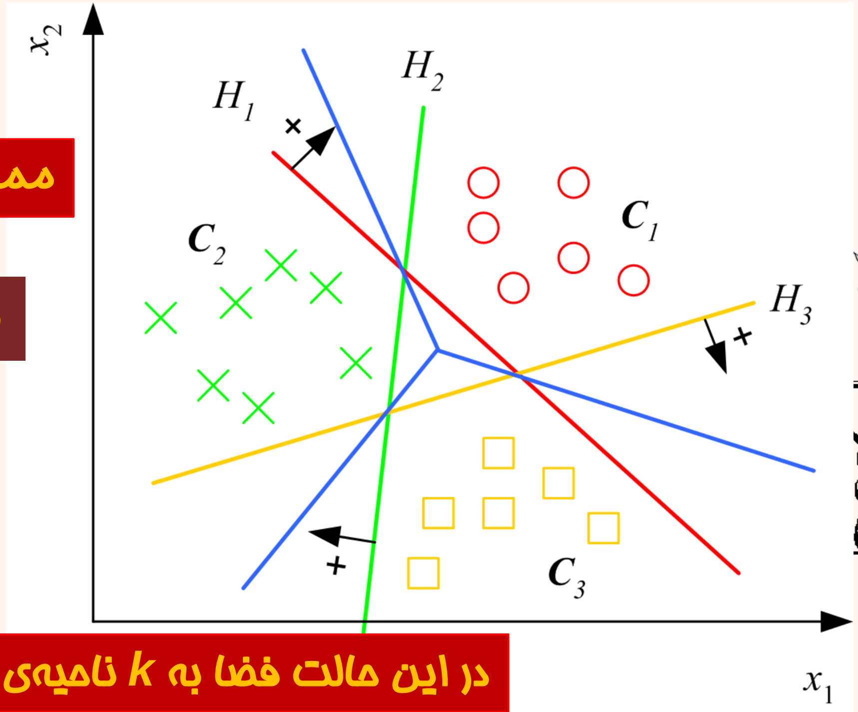
- در این حالت به k تابع جداساز نیاز است:

$$g_i(\mathbf{x}|\mathbf{w}_i, w_{i0}) = \mathbf{w}_i^T\mathbf{x} + w_{i0}$$

در نظر گرفتن چنین جداسازی به معنای این است که همه‌ی دسته‌ها جدایی‌پذیر خطی در نظر گرفته شوند.

$$g_i(\boldsymbol{x}|\boldsymbol{w}_i, w_{i0}) = \begin{cases} > 0 & \text{if } \boldsymbol{x} \in C_i \\ \leq 0 & \textit{otherwise} \end{cases}$$

Linearly separable



ممکن است همه‌ی دسته‌ها جدایی‌پذیر خطی نباشند:

Choose C_i if

Linear machine

$$g_i(\mathbf{x}) = \max_{j=1}^{K} g_j(\mathbf{x})$$

در این حالت فضا به k ناحیه‌ی محدب تقسیم می‌شود.

جداساز دوبه‌دو

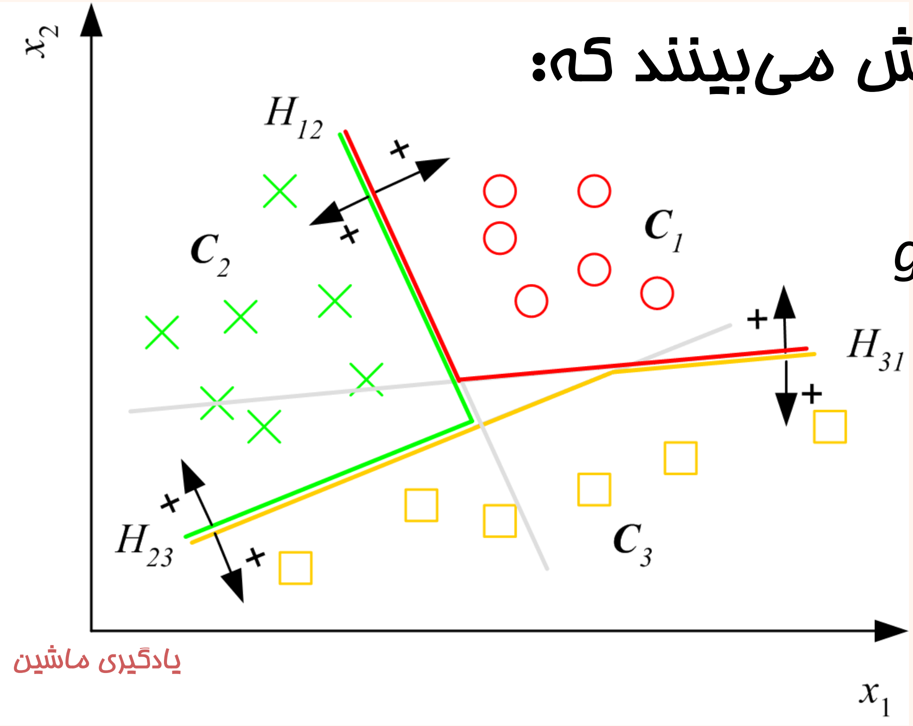
Pairwise Separation

one-versus-one

- اگر همه‌ی کلاس‌ها جدایی‌پذیر خطی نباشند، یک رویکرد مناسب تقسیم مسأله به چند جدایی‌ساز خطی است.
  - برای هر دو کلاس یک جداساز تعریف شود.

در این صورت $k(k-1)/2$ جداساز مورد نیاز است.

$$g_{ij}\left(\mathbf{x} \mid \mathbf{w}_{ij}, w_{ij0}\right) = \mathbf{w}_{ij}^T \mathbf{x} + w_{ij0}$$



- پارامترها به گونه‌ای آموزش می‌بینند که:

$$g_{ij}(\mathbf{x}) = \begin{cases} > 0 & \text{if } \mathbf{x} \in C_i \\ \leq 0 & \text{if } \mathbf{x} \in C_j \\ \text{don't care} & \text{otherwise} \end{cases}$$

- در زمان آزمایش:

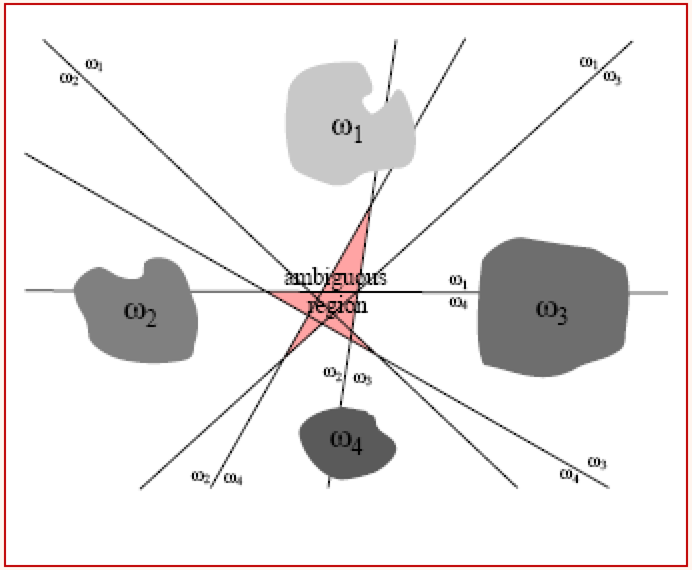
choose C_i if

$$\forall j \neq i, g_{ij}(\mathbf{x}) > 0$$

جداساز دوبه‌دو

- در این حالت ممکن است برخی نواحی رابطه‌ی پیش برقرار نباشد.
- در این حالت معیار دیگری را می‌توان جایگزین نمود:

$$g_i(\boldsymbol{x}) = \sum_{j \neq i} g_{ij}(\boldsymbol{x})$$



Duda et. al.

این شیوه نیز نمونه‌ی دیگری از تبدیل یک مسأله‌ی پیچیده به چند مسأله‌ی ساده‌تر است

Gradient-Descent

نزول گرادیان

- در دسته‌بندی مبتنی بر جداساز، پارامترها به گونه‌ای بهینه‌سازی می‌شوند که خطای دسته‌بندی حداقل شود:

$\boldsymbol{w}^*=\arg\min_{\boldsymbol{w}} E(\boldsymbol{w} \mid X)$

- در بیشتر مواقع، راه حل تحلیلی برای یافتن پارامترها وجود ندارد و چاره‌ای جز استفاده از یک روش بهینه‌سازی تکرارشونده نخواهد بود.
- استفاده از نزول گرادیان یکی از پرکاربردترین راه‌کارهاست.
- بردار گردایان به صورت زیر تعریف می‌شود:

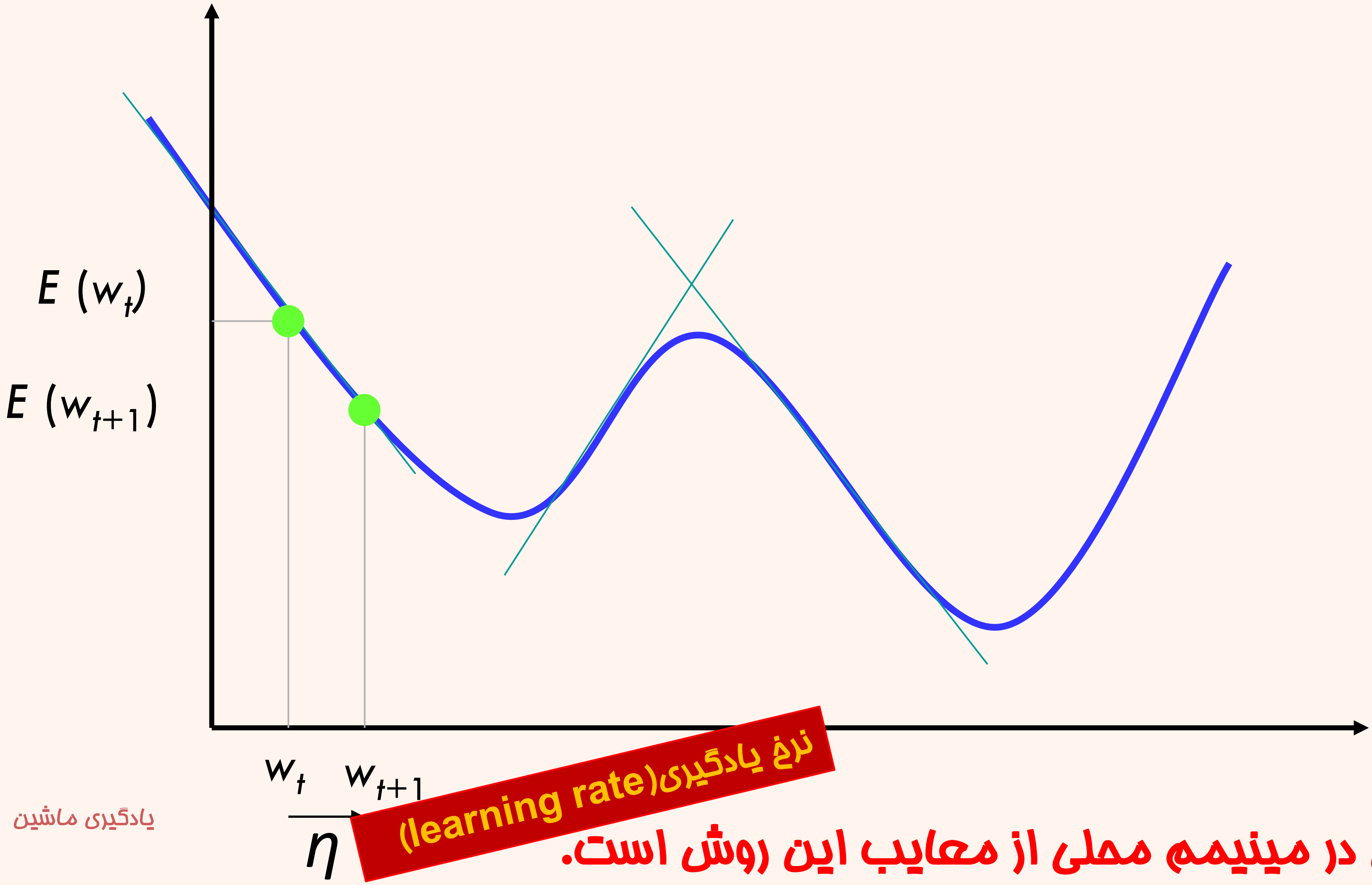
$$\nabla_w E = \left[\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \ldots, \frac{\partial E}{\partial w_d}\right]^T$$

Gradient-Descent

$$\Delta w_t = -\eta \frac{\partial E}{\partial w_t}, \forall i$$

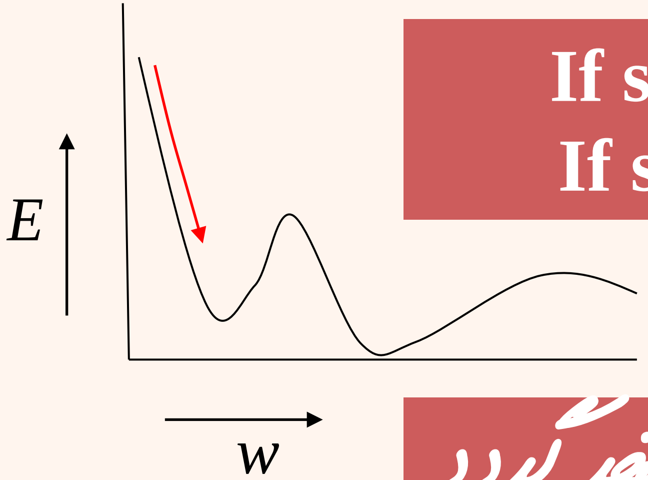
$$w_{t+1} = w_t + \Delta w_t$$

- پارامترها با یک مقدار تصادفی مقداردهی می‌شوند.
- بر خلاف جهت گرادیان مقدار پارامترها را به روز می‌شوند.



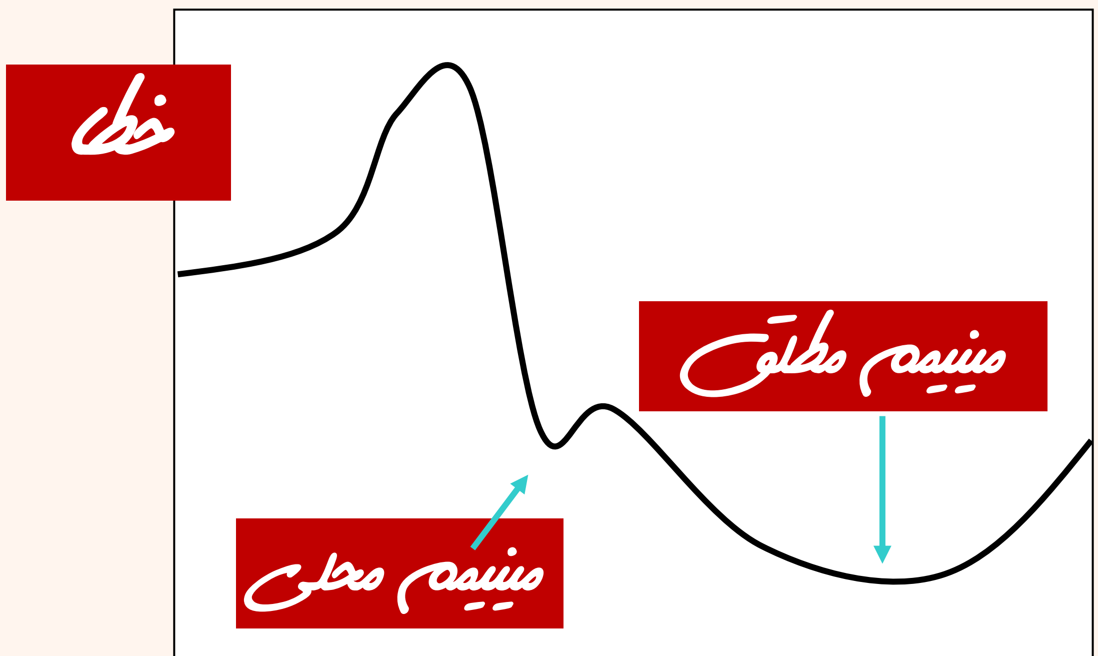
گیر کردن در مینیمم محلی از معایب این روش است.

Gradient Descent



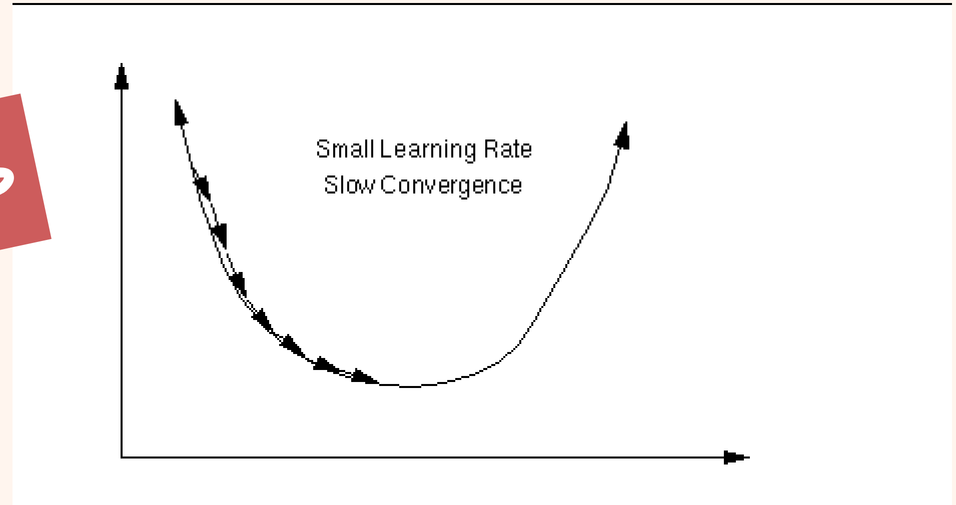
If slope is negative → increase w
If slope is positive → decrease w

مینیمم محلی جایی است که مشتق صفر گردد

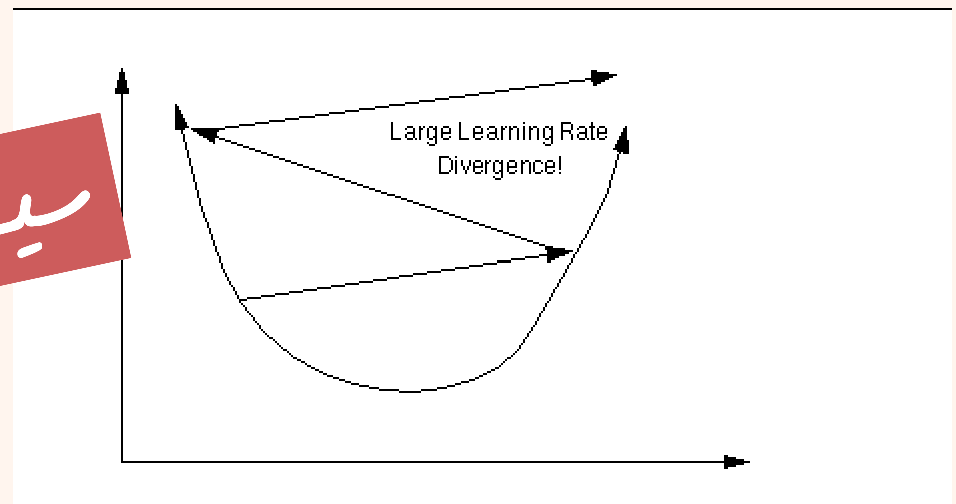


تنظیم نرخ یادگیری

همگرایی کند است.

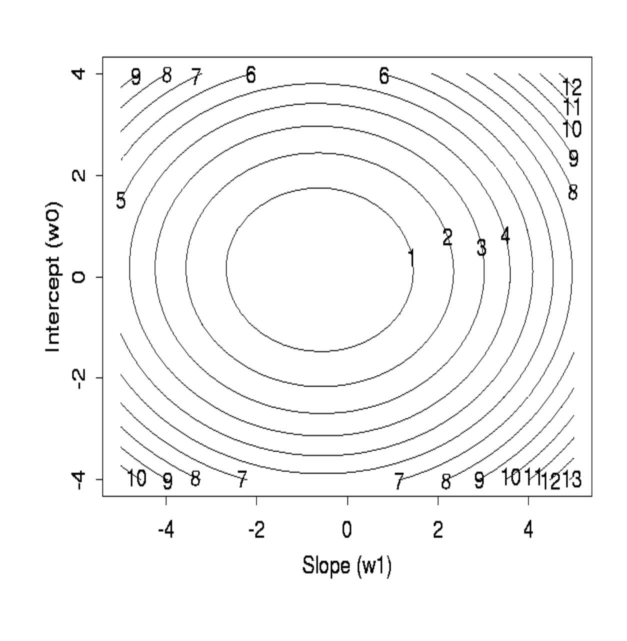


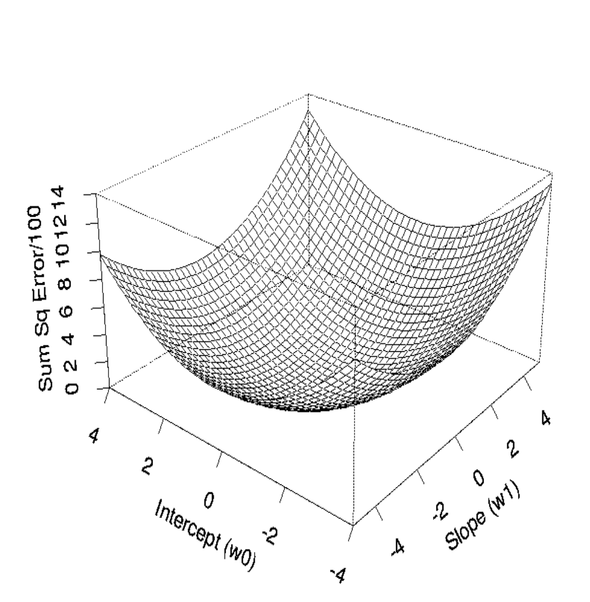
سیستم ناپایدار است.

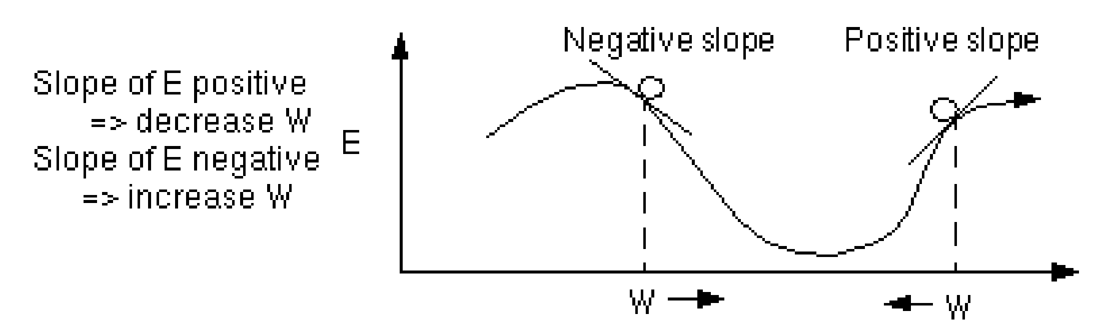


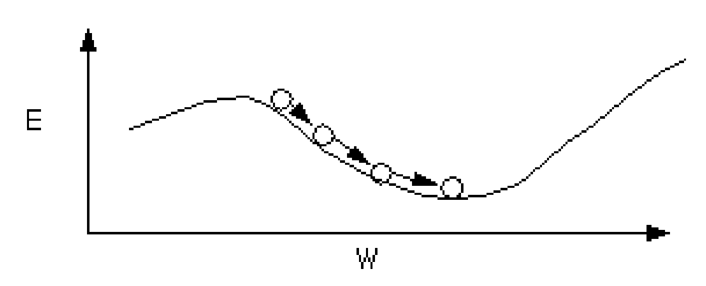
کمینه کردن خطا(ادامه...)

Steepest descent rule









انواع شیوه‌های آموزش

- آموزش به دو صورت قابل انجام است:
- دسته‌ای:

در این شیوه تابع خطایی که بناست مورد استفاده قرار گیرد، تابع خطا به ازای اعمال همه‌ی ورودی‌هاست.

- ترتیبی:

در این شیوه ورودی‌های تک‌تک اعمال شده، پس از محاسبه‌ی خطا، پارامترها اصلاح شده و به همین ترتیب ورودی بعدی اعمال می‌شود.

مروری بر جداسازی پارامتری

- در صورتی که داده‌ها از توزیع گاوسی تبعیت کنند و ماتریس کواریانس یکسانی داشته باشند، جداساز بهینه، خطی خواهد بود:

$$p(\boldsymbol{x} \mid C_i) \sim N(\boldsymbol{\mu}_i, \Sigma)$$

$$g_i(\mathbf{x} \mid \mathbf{w}_i, w_{i0}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

$$\mathbf{w}_i = \Sigma^{-1}\mu_i \quad w_{i0} = -\frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i + \log P(C_i)$$

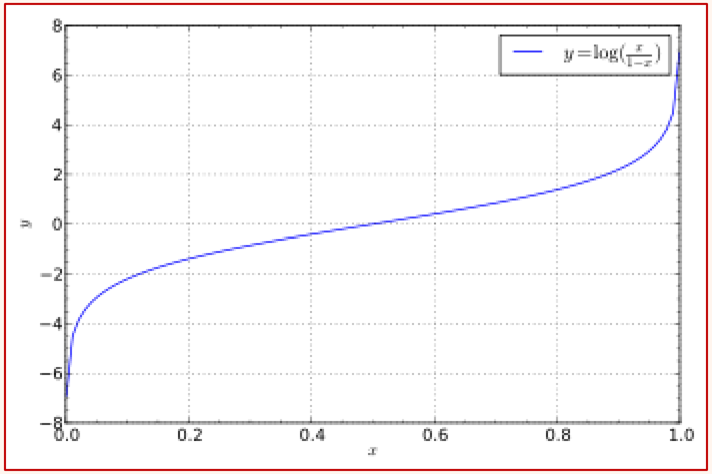
- برای یک مجموعه‌ی آموزشی ابتدا مقادیر میانگین و ماتریس کواریانس هر کلاس محاسبه می‌شود.
- برای حالت دو کلاسه خواهیم داشت:

$$y \equiv P(C_1 \mid \boldsymbol{x}) \text{ and } P(C_2 \mid \boldsymbol{x}) = 1 - y$$

$$\text{choose } C_1 \text{ if } \begin{cases} y > 0.5 \\ y/(1-y) > 1 \\ \log\left[y/(1-y)\right] > 0 \end{cases} \text{ and } C_2 \text{ otherwise}$$

مروری بر جداسازی پارامتری(ادامه...)

- LOGIT به صورت زیر تعریف می‌شود:



$$\operatorname{logit}(y) = \log \frac{y}{1-y}$$

در نتیجه داده به کلاس یک تعلق دارد اگر و تنها اگر

$$\operatorname{logit}(P(C_1|x)) > 0$$

وگرنه متعلق به کلاس دو خواهد بود.

مروری بر جداسازی پارامتری(ادامه...)

$$\text{logit}\left(P\left(C_1|\boldsymbol{x}\right)\right)=\log\frac{P\left(C_1|\boldsymbol{x}\right)}{1-P\left(C_1|\boldsymbol{x}\right)}=\log\frac{P\left(C_1|\boldsymbol{x}\right)}{P\left(C_2|\boldsymbol{x}\right)}$$

$$=\log\frac{p\left(\boldsymbol{x}|C_1\right)}{p\left(\boldsymbol{x}|C_2\right)}+\log\frac{P\left(C_1\right)}{P\left(C_2\right)}$$

$$=\log\frac{(2\pi)^{-d/2}|\Sigma|^{-1/2}\exp\left[-(1/2)(\boldsymbol{x}-\mu_1)^T\Sigma^{-1}(\boldsymbol{x}-\mu_1)\right]}{(2\pi)^{-d/2}|\Sigma|^{-1/2}\exp\left[-(1/2)(\boldsymbol{x}-\mu_2)^T\Sigma^{-1}(\boldsymbol{x}-\mu_2)\right]}+\log\frac{P\left(C_1\right)}{P\left(C_2\right)}$$

$$=\boldsymbol{w}^T\boldsymbol{x}+w_0$$

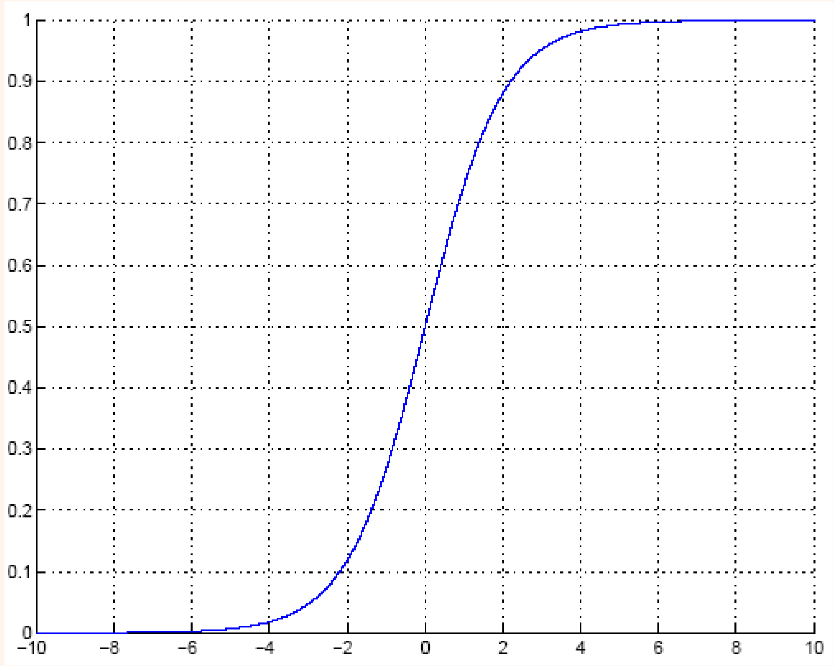
$$\text{where } \boldsymbol{w}=\Sigma^{-1}(\mu_1-\mu_2)\quad w_0=-\frac{1}{2}(\mu_1+\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)+\log\frac{P\left(C_1\right)}{P\left(C_2\right)}$$

The inverse of logit

$$\log\frac{P\left(C_1|\boldsymbol{x}\right)}{1-P\left(C_1|\boldsymbol{x}\right)}=\boldsymbol{w}^T\boldsymbol{x}+w_0$$

$$P\left(C_1|\boldsymbol{x}\right)=\frac{1}{1+\exp\left[-\left(\boldsymbol{w}^T\boldsymbol{x}+w_0\right)\right]}\quad =\text{sigmoid}\left(\boldsymbol{w}^T\boldsymbol{x}+w_0\right)$$

تابع sigmoid



Calculate $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ and choose C_1 if $g(\mathbf{x}) > 0$, or

Calculate $y = \text{sigmoid}\left(\mathbf{w}^T\mathbf{x} + w_0\right)$ and choose C_1 if $y > 0.5$

در حالت دوم تابع sigmoid ارتباط مقدار تابع جداساز و احتمال پسین را نشان می‌دهد.

Logistic Discrimination

حالت دوکلاسه

Logistic Regression

- در دسته‌بندی مبتنی بر درست‌نمایی ابتدا، $P(C_1)$ و $P(x|C_1)$ محاسبه شده، سپس مقدار $P(C_1|x)$ به دست می‌آید.
- در logistic discrimination احتمال پسین به صورت مستقیم برآورد می‌شود.
- در صورتی که لگاریتم نسبت درست‌نمایی دو کلاس خطی باشد:

$$\log \frac{p(\boldsymbol{x}|C_1)}{p(\boldsymbol{x}|C_2)} = \boldsymbol{w}^T \boldsymbol{x} + w_0^o$$

- با توجه به قانون Bayes:

$$\operatorname{logit}\left(P(C_1|\boldsymbol{x})\right) = \log \frac{P(C_1|\boldsymbol{x})}{1 - P(C_1|\boldsymbol{x})} = \log \frac{p(\boldsymbol{x}|C_1)}{p(\boldsymbol{x}|C_2)} + \log \frac{P(C_1)}{P(C_2)}$$

حالت دوکلاسه

$$\operatorname{logit}\left(P\left(C_1|\boldsymbol{x}\right)\right) = \log \frac{P\left(C_1|\boldsymbol{x}\right)}{1-P\left(C_1|\boldsymbol{x}\right)} = \log \frac{p\left(\boldsymbol{x}|C_1\right)}{p\left(\boldsymbol{x}|C_2\right)} + \log \frac{P\left(C_1\right)}{P\left(C_2\right)}$$

$$= \boldsymbol{w}^T \boldsymbol{x} + w_0$$

$$\text{where } w_0 = w_0^o + \log \frac{P\left(C_1\right)}{P\left(C_2\right)}$$

$$y = \hat{P}\left(C_1|\boldsymbol{x}\right) = \frac{1}{1+\exp\left[-\left(\boldsymbol{w}^T \boldsymbol{x} + w_0\right)\right]}$$

بدین‌ترتیب، تقریبی از احتمال پسین به دست می‌آید.

در واقع مسأله یافتن (آموختن) w و w_0 است.

آموزش (حالت دو کلاسه)

$$y = \hat{P}(C_1|\boldsymbol{x}) = \frac{1}{1+\exp\left[-\left(\boldsymbol{w}^T\boldsymbol{x}+w_0\right)\right]}$$

$$\mathcal{X} = \{\boldsymbol{x}^t, r^t\}_t \qquad r^t|\boldsymbol{x}^t \sim \text{Bernoulli}\left(y^t\right)$$

در این جا به صورت مستقیم، مقدار y تخمین زده می‌شود.
درست‌نمایی $\mathcal{X}$ به ازای پارامترهای مورد نظر محاسبه شده:

$$l(\boldsymbol{w}, w_0|\mathcal{X}) = \prod_t \left(y^t\right)^{\left(r^t\right)}\left(1-y^t\right)^{\left(1-r^t\right)}$$

بر اساس آن تابع خطا محاسبه می‌شود:

$$E = -\log l$$

Binary cross entropy

$$E(\boldsymbol{w}, w_0|\mathcal{X}) = -\sum_t r^t \log y^t + \left(1-r^t\right)\log\left(1-y^t\right)$$

هدف کاهش میزان خطا است! هیچ راه تحلیلی برای حل این مسأله وجود ندارد.

آموزش(حالت دو کلاسه)، نزول گرادیان

$$y = \hat{P}(C_1|\boldsymbol{x}) = \frac{1}{1+\exp\left[-\boxed{\left(\boldsymbol{w}^T\boldsymbol{x}+w_0\right)}\right]}$$

g(x)

$$E(\boldsymbol{w}, w_0|\mathcal{X}) = -\sum_t r^t \log y^t + \left(1-r^t\right)\log\left(1-y^t\right)$$

$$\text{If } y = \text{sigmoid}(a) \qquad \frac{dy}{da} = y(1-y)$$

$$\Delta w_j = -\eta\frac{\partial E}{\partial w_j}$$

$$\frac{\partial E}{\partial w_j} = \boxed{\frac{\partial E}{\partial y^t}} \times \boxed{\frac{\partial y^t}{\partial g}} \times \boxed{\frac{\partial g}{\partial w_j}}$$

Chain rule

$$= \eta\sum_t \boxed{\left(\frac{r^t}{y^t} - \frac{1-r^t}{1-y^t}\right)} \boxed{y^t\left(1-y^t\right)} \boxed{x_j^t}$$

$$= \eta\sum_t \left(r^t - y^t\right)x_j^t,\ j = 1,\ldots,d$$

$$\Delta w_0 = -\eta\frac{\partial E}{\partial w_0} = \eta\sum_t\left(r^t - y^t\right)$$

الگوریتم

For $j = 0, \ldots, d$

$\quad w_j \leftarrow$rand(-0.01,0.01)

Repeat

$\quad$For $j = 0, \ldots, d$

$\quad\quad \Delta w_j \leftarrow 0$

$\quad$For $t = 1, \ldots, N$

$\quad\quad o \leftarrow 0$

$\quad\quad$For $j = 0, \ldots, d$

$\quad\quad\quad o \leftarrow o + w_j x_j^t$

$\quad\quad y \leftarrow \text{sigmoid}(o)$

$\quad\quad \Delta w_j \leftarrow \Delta w_j + (r^t - y) x_j^t$

$\quad$For $j = 0, \ldots, d$

$\quad\quad w_j \leftarrow w_j + \eta \Delta w_j$

Until convergence

Regularization

منظم‌سازی

$$E\left(\boldsymbol{w}, w_0|\mathcal{X}\right) = -\sum_t r^t \log y^t + \left(1-r^t\right)\log\left(1-y^t\right) + \frac{\lambda}{2}\|W\|^2$$

K>2 Classes

حالت چنددسته‌ای

$$\mathcal{X} = \left\{ \boldsymbol{x}^t, \boldsymbol{r}^t \right\}_t \quad r^t | \boldsymbol{x}^t \sim \text{Mult}_K\left(1, \boldsymbol{y}^t\right)$$

$$\log \frac{p\left(\boldsymbol{x}|C_i\right)}{p\left(\boldsymbol{x}|C_K\right)} = \boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}^o$$

$$\frac{p\left(C_i|\boldsymbol{x}\right)}{p\left(C_K|\boldsymbol{x}\right)} = \exp\left[\boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}\right] \qquad w_{i0} = w_{i0}^o + \log P\left(C_i\right)/P\left(C_k\right)$$

$$\sum_{i=1}^{K-1} \frac{p\left(C_i|\boldsymbol{x}\right)}{p\left(C_K|\boldsymbol{x}\right)} = \frac{1 - p\left(C_K|\boldsymbol{x}\right)}{p\left(C_K|\boldsymbol{x}\right)} = \sum_{i=1}^{K-1} \exp\left[\boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}\right]$$

$$p\left(C_K|\boldsymbol{x}\right) = \frac{1}{1 + \sum_{i=1}^{K-1} \exp\left[\boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}\right]}$$

حالت چنددسته‌ای

$$p\left(C_K|\boldsymbol{x}\right)=\frac{1}{1+\sum_{i=1}^{K-1}\exp\left[\boldsymbol{w}_i^T\boldsymbol{x}+w_{i0}\right]}$$

$$\frac{p\left(C_i|\boldsymbol{x}\right)}{p\left(C_K|\boldsymbol{x}\right)}=\exp\left[\boldsymbol{w}_i^T\boldsymbol{x}+w_{i0}\right]$$

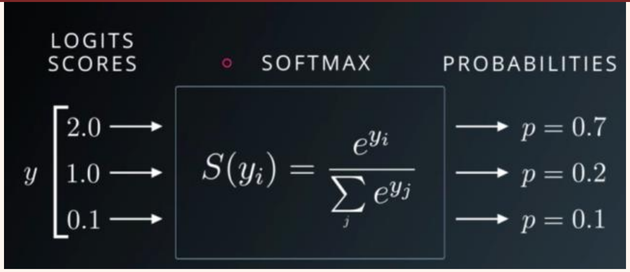
$$p\left(C_i|\boldsymbol{x}\right)=\frac{\exp\left[\boldsymbol{w}_i^T\boldsymbol{x}+w_{i0}\right]}{1+\sum_{i=1}^{K-1}\exp\left[\boldsymbol{w}_i^T\boldsymbol{x}+w_{i0}\right]}$$

برای این که با همه‌ی کلاس‌ها یکسان برخورد شود:

$$y_i=\hat{P}\left(C_i|\boldsymbol{x}\right)=\frac{\exp\left[\boldsymbol{w}_i^T\boldsymbol{x}+w_{i0}\right]}{\sum_{j=1}^{K}\exp\left[\boldsymbol{w}_j^T\boldsymbol{x}+w_{j0}\right]},\, i=1,\ldots,K$$

softmax

محاسبه‌ی ماکزیمم است، با این تفاوت که مشتق‌پذیر است.



K>2 Classes

حالت چنددسته‌ای

$$\mathcal{X} = \{\boldsymbol{x}^t, \boldsymbol{r}^t\}_t \quad r^t | \boldsymbol{x}^t \sim \mathrm{Mult}_K(1, \boldsymbol{y}^t)$$

$$l\left(\{\boldsymbol{w}_i, w_{i0}\}_i | \mathcal{X}\right) = \prod_t \prod_i \left(y_i^t\right)^{\left(r_i^t\right)}$$

cross entropy

$$E\left(\{\boldsymbol{w}_i, w_{i0}\}_i | \mathcal{X}\right) = -\sum_t \sum_i r_i^t \log y_i^t$$

هدف کاهش میزان خطا است. هیچ راه تحلیلی برای حل این مسأله وجود ندارد.

$$\Delta \boldsymbol{w}_j = \eta \sum_t \left(r_j^t - y_j^t\right) \boldsymbol{x}^t \quad \Delta w_{j0} = \eta \sum_t \left(r_j^t - y_j^t\right)$$

الگوریتم

For $i = 1, \ldots, K$, For $j = 0, \ldots, d$, $w_{ij} \leftarrow \text{rand}(-0.01, 0.01)$

Repeat

  For $i = 1, \ldots, K$, For $j = 0, \ldots, d$, $\Delta w_{ij} \leftarrow 0$

  For $t = 1, \ldots, N$

    For $i = 1, \ldots, K$

      $o_i \leftarrow 0$

      For $j = 0, \ldots, d$

        $o_i \leftarrow o_i + w_{ij} x_j^t$

    For $i = 1, \ldots, K$

      $y_i \leftarrow \exp(o_i) / \sum_k \exp(o_k)$

    For $i = 1, \ldots, K$

      For $j = 0, \ldots, d$

        $\Delta w_{ij} \leftarrow \Delta w_{ij} + (r_i^t - y_i) x_j^t$

  For $i = 1, \ldots, K$

    For $j = 0, \ldots, d$

      $w_{ij} \leftarrow w_{ij} + \eta \Delta w_{ij}$

Until convergence