均方根差：MSE $E(y-\hat{f}(x))^2$

方差：Var $Var(\hat{f}(x)) = E[\hat{f}(x)^2] - E[\hat{f}(x)]^2$

偏差：Bias $Bias(\hat{f}(x)) = E(\hat{f}(x)) - f(x)$

$Var[X] = E[X^2] - (E[X])^2$

$$Var[X] \frac{\sum_{i=1}^{m}(X-\bar{X})^2}{m} = \frac{1}{m}\sum_{i=1}^{m}(X^2+\bar{X}^2-2X\bar{X})$$

$$= \frac{\sum_{i=1}^{m}X^2}{m} + \frac{\sum_{i=1}^{m}\bar{X}^2}{m} - \frac{\sum_{i=1}^{m}2X\bar{X}}{m}$$

$$E(X^2) + \bar{X}^2 - 2\bar{X}E(X)$$

$$= E(X^2) + (E(X))^2 - 2(E(X))^2$$

$$= E(X^2) - (E(X))^2$$

$y = f + \varepsilon$, ε (noise, zero mean and variance σ^2)

$E[(y-\hat{f})^2] = MSE$

$$= E[(f+\varepsilon-\hat{f})^2]$$

$$= E[(f+\varepsilon-\hat{f}+E(\hat{f})-E(\hat{f}))^2]$$

$$= E[(f-E(\hat{f}))^2] + E[(E(\hat{f})-\hat{f})^2]$$

$$+ E(\varepsilon^2) + 2E[(f-E(\hat{f}))\cdot(E(\hat{f})-\hat{f})]$$

$$+ 2E[(f-E(\hat{f}))\varepsilon] + 2E[(E(\hat{f})-\hat{f})\cdot\varepsilon]$$

$$= (f-E(\hat{f}))^2 + E(\varepsilon^2) + E[(E(\hat{f})-\hat{f})^2]$$

$$= (f-E(\hat{f}))^2 + Var(\varepsilon) + Var(\hat{f})$$

$$= Bias(\hat{f})^2 + \sigma^2 + Var(\hat{f})$$

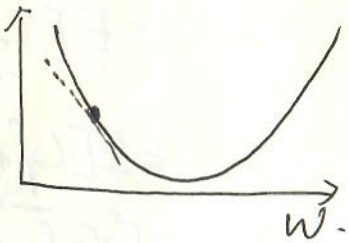
梯度下降：

$J(w,b)=\frac{1}{m}\sum_{i=1}^{m} J(\hat{y}^{(i)},y^{(i)})$

find w and b to minimize $J(w,b)$

$$w := w - \alpha\frac{dJ(w,b)}{dw}$$

$$b := b - \alpha\frac{dJ(w,b)}{db}$$

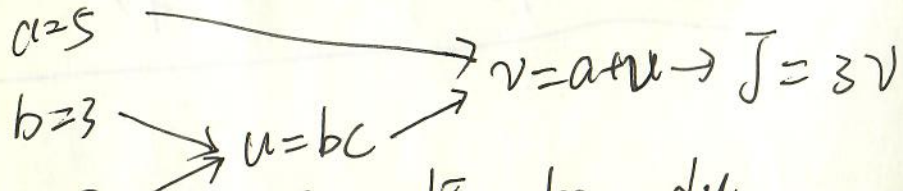


2.5 导数

$$e^x dx = e^x$$

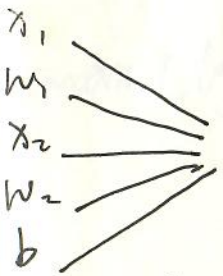
2.6 计算图

$\frac{d\,\text{Final Output}}{d\,\text{Var}}$



2.9 logistic 回归中的梯度下降

从最后对前一项依次求导，得到结果对输入的导数。



$z=w_1x_1+w_2x_2+b \rightarrow a=\sigma(z) \rightarrow L(a,y)$

$$da=\frac{dL(a,y)}{da}=\frac{d\left(-[y\log a+(1-y)\log(1-a)]\right)}{da}$$

反函数的导数等于直接函数导数的倒数。

$$=-\left(\frac{y}{a\ln e}+1\cdot\frac{1-y}{(1-a)\ln e}\right)$$

$$=-\frac{y}{a}+\frac{1-y}{1-a}$$

$(a=\hat{y})$

$$dw_1=\frac{dL(a,y)}{dw_1}=\frac{dL(a,y)}{da}\cdot\frac{da}{dz}\cdot\frac{dz}{dw_1}$$

$$=\left(-\frac{y}{a}+\frac{1-y}{1-a}\right)\cdot a(1-a)\cdot x_1$$

2.11 向量化 vectorization.

$$X=\begin{bmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \cdots & x^{(m)} \\ | & | & & | \end{bmatrix} \quad Y=[y^{(1)}, y^{(2)} \cdots y^{(m)}]$$

$$Z=[z^{(1)}\ z^{(2)} \cdots z^{(m)}]=w^T X+[b\ b \cdots b]$$

$$Z=np.dot(w^T, X)+b$$

$$A=[a^{(1)} \cdots a^{(m)}] \quad Y=[y^{(1)} \cdots y^{(m)}]$$

$$dZ=A-Y$$

2.15 Broadcasting

use np.random.randn(1,5)

but not np.random.randn(5)

Can use reshape to set the dim of the matrix.

2.17 logistic cost function

If $y=1$: $P(y|x)=\hat{y}$

If $y=0$: $P(y|x)=1-\hat{y}$

$$P(y|x)=\hat{y}^y(1-\hat{y})^{(1-y)}$$

$$\max(\log(P(y|x)))=\log(\hat{y}^y(1-\hat{y})^{(1-y)})$$

$$=y\log\hat{y}+(1-y)\log(1-\hat{y})$$

$$=-L(y,\hat{y})$$

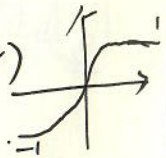
$$\min(-\log(P(y|x)))=-(y\log\hat{y}+(1-y)\log(1-\hat{y}))$$

$$=L(y,\hat{y})$$

3.6 激活函数

sigmoid function: $a=\frac{1}{1+e^{-z}}$

$a=\tanh(z)=\frac{e^z-e^{-z}}{e^z+e^{-z}}$ (better)



ReLu: $a=\max(0,z)$

3.10

$$\frac{dL}{dz^{[2]}} = \frac{dL}{da^{[2]}} \frac{da^{[2]}}{dz^{[2]}} = a^{[2]} - y$$

$$\frac{dL}{dw^{[2]}} = \frac{dL}{dz^{[2]}} \cdot \frac{dz^{[2]}}{dw^{[2]}} = (a^{[2]} - y) \cdot a^{[1]T}$$

$$\frac{dL}{db} = \frac{dL}{dz^{[2]}} \cdot 1 = dz^{[2]}$$

$$\frac{dL}{dz^{[1]}} = \frac{dL}{dz^{[2]}} \cdot \frac{dz^{[2]}}{dz^{[1]}}$$

$$= \frac{dL}{dz^{[2]}} \cdot \frac{dz^{[2]}}{da^{[1]}} \cdot \frac{da^{[1]}}{dz^{[1]}}$$

$$= \frac{dL}{dz^{[2]}} \cdot w^{[2]} \cdot$$

$$= w^{[2]T} dz^{[2]} \cdot g^{[1]'}(z^{[1]})$$

$$dw^{[1]} = dz^{[1]} x^T$$

$$db^{[1]} = dz^{[1]}$$

2019 of 24.

为什么无偏估计的方差是除以 $n-1$?

$$\delta^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$E(XY) = EX \cdot EY$

$Var(\bar{X}) = \frac{1}{n} Var(X)$?

$\delta(\bar{x})^2 = \frac{1}{n} \delta(x)^2$

二阶中心距:

$$Var(X) = E(X^2) - (E(X))^2$$

$$E(X^2) = Var(X) + (E(X))^2 = \delta^2 + \mu^2$$

$$E(\bar{X}^2) = Var(\bar{x}) + (E(\bar{x}))^2 = \frac{1}{n}\delta^2 + \mu^2$$

无偏估计: 对变量 θ 的估计是 $\hat{\theta}$, 如果 $E(\hat{\theta}) = E(\theta)$, 则称 $\hat{\theta}$ 为 θ 的无偏估计.

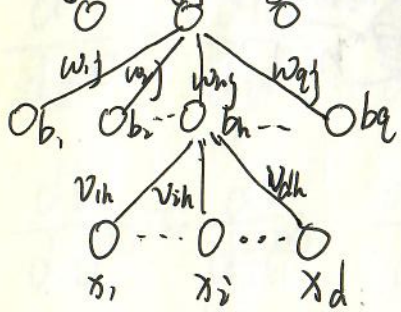
$$E\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)$$

$$= \sum_{i=1}^{n} E(x_i^2) - n \cdot E(\bar{x}^2)$$

$$= \sum_{i=1}^{n} (\delta^2 + \mu^2) - n\left(\frac{1}{n}\delta^2 + \mu^2\right)$$

$$= (n-1)\delta^2$$

2019/10/02 单个样本 BP (BackPropagation)



输出层. 第j个输出神经元的输入 $\beta_j = \sum_{h=1}^{q} w_{hj} b_h$

隐层 第h个隐层神经元的输入 $\alpha_h = \sum_{i=1}^{d} v_{ih} x_i$

输入层 (x_k, y_k)

输出: $\hat{y}_k = (\hat{y}_1^k, \hat{y}_2^k, \cdots \hat{y}_l^k)$

$\hat{y}_j^k = f(\beta_j - \theta_j)$ (sigmoid) $\quad \beta_j = \sum_{h=1}^{q} w_{hj} b_h$

$\alpha_h = \sum_{i=1}^{d} v_{ih} x_i$

网络在(x_k, y_k)的均方误差:

$E_k = \frac{1}{2}\sum_{j=1}^{l} (\hat{y}_j^k - y_j^k)^2$

[以定网络可以表示映射关系的前提下, 计算误差来优化参数]

需确定的参数有:

$d \times q + q \times l + q + l$

输入 输出 隐层 输出层
权值 权值 阈值 阈值

使梯度下降: $v \leftarrow v + \Delta v$

更新参数: $\Delta w_{hj} = -\eta \frac{\partial E_k}{\partial w_{hj}}$

(单个训练样本更新)

误差函数对权重求导使偏导为0

$$\frac{\partial E_k}{\partial w_{hj}} = \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} \cdot \frac{\partial \beta_j}{\partial w_{hj}}$$

$$\frac{\partial \beta_j}{\partial w_{hj}} = b_h$$

$\frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j}$ ② 由于 $\hat{y}_j^k = f(\beta_j - \theta_j)$ 是sigmoid函数

所以 $\frac{\partial \hat{y}_j^k}{\partial \beta_j} = (\hat{y}_j^k)' = f(x) \cdot (1 - f(x))$

$$g_j = -\frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} = -(\hat{y}_j^k - y_j^k) \cdot \hat{y}_j^k \cdot (1 - \hat{y}_j^k)$$

$$= \hat{y}_j^k \cdot (1 - \hat{y}_j^k) \cdot (y_j^k - \hat{y}_j^k)$$

(学习率)

(权重) $\Delta w_{hj} = \eta g_j b_h$

$\Delta \theta_j = -\eta g_j$ (阈值) $\quad \Delta \gamma_h = -\eta e_h$ (阈值)

(权重) $\Delta v_{ih} = -\frac{\partial E_k}{\partial b_h} \cdot \frac{\partial b_h}{\partial \alpha_h} \cdot \frac{\partial \alpha_h}{\partial v_{ih}} \cdot \eta$

$= -\sum_{j=1}^{l} \frac{\partial E_k}{\partial \beta_j} \cdot \frac{\partial \beta_j}{\partial b_h} \cdot f'(\alpha_h - \gamma_h) \cdot \eta$

$= -\sum_{j=1}^{l} -g_j w_{hj} \cdot b_h (1 - b_h) = b_h (1 - b_h) \sum_{j=1}^{l} g_j w_{hj}$

累积BP

将第个 E_k 误差相加，则对 w_{hj} 求偏导也是各项的偏导相加.

$$w_{hj} \leftarrow w_{hj} + \frac{1}{n}\Delta w_{hj}$$

方差与偏差的平衡：使用正则化减小方差，避免过拟合 (overfitting).

$$E = \lambda \frac{1}{m}\sum_{k=1}^{m} E_k + (1-\lambda)\sum \frac{1}{2} w_i^2$$

(连接权和阈值，也就是也可用 n^2 参数值. 参数)

BP算法会导到局部最小，为了得到比某局部最小更优的解，可以：

① 以不同的参数值初始化多个神经网络

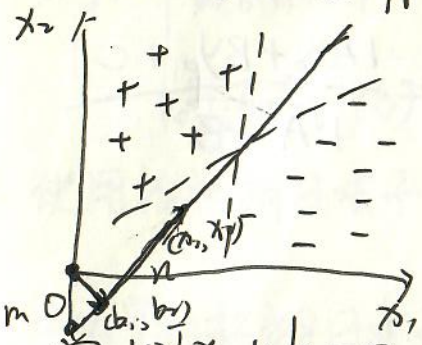
② "模拟退火" 以一定的概率接受比当前解更差的结果 (保证每个最小参数空间，规定迭代次数)

③ 随机梯度下降，即便在局部最小点，它计算的梯度仍可能不为零.

增加隐层数目提高拟合能力.

2019.10.04

I 支持向量机 (Support Vector machine)



分割两类点的超平面

$$w^T x + b = 0$$

dimension

$$w = (w_1; w_2; w_3; \cdots; w_d)$$

当 $w^T x + b > 0$，点在 $w^T x + b$ 的上方

当 $w^T x + b < 0$，点在 $w^T x + b$ 的下方.

w 为超平面 $w^T x + b = 0$ 的法向量

由于向量 x 可以分解为 m 与 n，$x = m + n$.

$$w^T(m+n) + b = 0$$

$$w^T m + w^T n + b = 0$$

由于 w 与 n 垂直，$w^T n = 0$

$$w^T m + b = 0, \quad w^T m = -b.$$

样本空间任意点 x 到超平面 (w, b) 距离：

$$r = \frac{|w^T x + b|}{\|w\|}$$

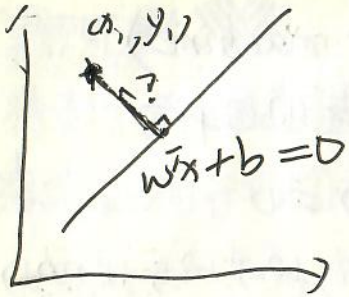
$[w_1, w_2]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + b$

$d = |\cos\theta|$

L_0 范数是指向量中非0的元素个数.

L_2 范数是指向量各元素平方和然后求平方根

L_1 范数是向量各元素之和



由于 $d=\frac{|Ax_0+By_0+C|}{\sqrt{A^2+B^2}}$

$$d=\frac{|w^Tx+b|}{\|w\|}=r.$$

$$\begin{cases} w^Tx_i+b\geq +1, & y_i=1 \\ w^Tx_i+b\leq -1, & y_i=-1 \end{cases}$$

两个异类支持向量到超平面的距离之和：

$r=\frac{2}{\|w\|}$，称为间隔.

这里先固定正样本点和负样本点到超平面 $w^Tx+b=0$ 的最小距离，再求最优的 w, b. 这里的问题是我们不一定能得到 1 和 -1 这样的最小距离，那我们可以设置得非常小 eg: 10^{-10}，只要大于 0，即可.

要 r 最大，

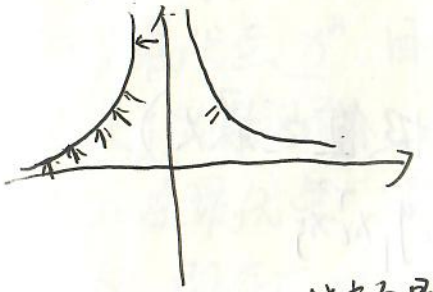
$$\begin{cases} \max\limits_{w,b} \frac{2}{\|w\|} \text{ (距离大)} \\ y_i(w^Tx_i+b)\geq 1 \text{ (分类正确)} \end{cases} \Rightarrow \begin{cases} \min\limits_{w,b} \|w\|^2 \\ y_i(w^Tx_i+b)\geq 1 \end{cases}$$

对于：$\begin{cases} \min\limits_{w,b} \|w\|^2 \text{ (凸二次规划问题)} \\ y_i(w^Tx_i+b)\geq 1, \ i=1,2,3,\cdots m \end{cases}$

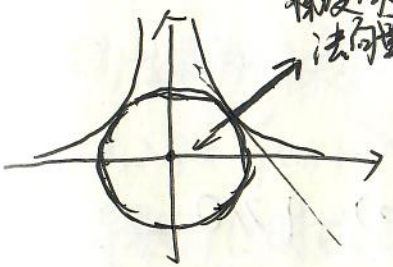
使用拉格朗日乘子法可得到其"对偶问题"

什么是拉格朗日乘子法？

$g(x,y)=x^2y$，其梯度向量 $\nabla g=\begin{pmatrix}\frac{\partial g}{\partial x}\\ \frac{\partial g}{\partial y}\end{pmatrix}=\begin{pmatrix}2xy\\ x^2\end{pmatrix}$



梯度向量是等高线的法向量，梯度向量与等高线的切线垂直.



$\begin{cases} \nabla f=\lambda\nabla g \text{ (确定梯度向量平行)} \\ x^2y=3 \text{ (等高线，确定梯度)} \end{cases}$

$g(x,y)=x^2y$ (到原点最小)

$f(x,y)=x^2+y^2$. 距离)

定义：函数 f 在 g 约束下的极值，这种问题可以表示为：

$\min\max\limits_{\lambda} f+\lambda g$

s.t. $\lambda\geq 0$

$\min f(x,y)=x^2+y^2$

subject to: $g(x,y)=xy-3=0$

对每条约束添加拉格朗日乘子 $\alpha_i \geq 0$.

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2+\sum_{i=1}^{m}\alpha_i(1-y_i(w^Tx_i+b))$$

$\alpha=(\alpha_1;\alpha_2;\cdots;\alpha_m)$

对 w 和 b 求偏导为零

$$\begin{cases} dw=w-\sum_{i=1}^{m}\alpha_i y_i x_i = w-\sum_{i=1}^{m}\alpha_i y_i x_i = 0 \\ db=\sum_{i=1}^{m}\alpha_i y_i = 0 \end{cases}$$

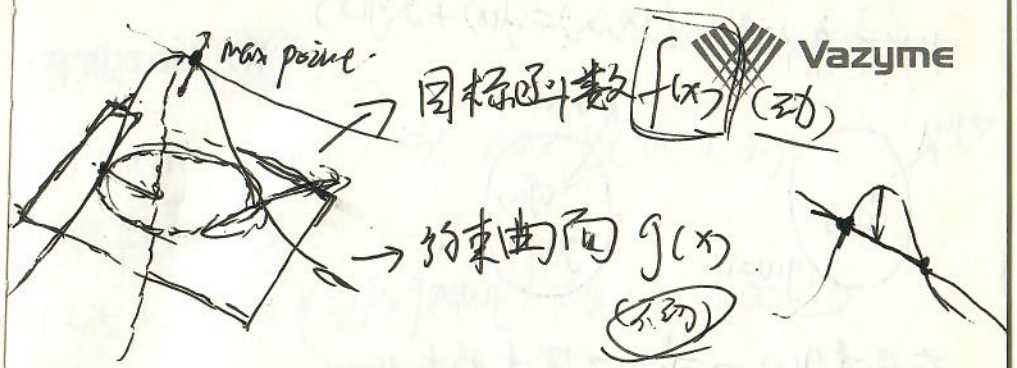
代入 $L(w,b,\alpha)$ 中：（极值点最大）

$$\begin{cases} \max\limits_{\alpha} \sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j \\ \quad s.t.\ \sum_{i=1}^{n}\alpha_i y_i=0 \end{cases}$$

$$\begin{cases} \alpha_i \geq 0 \\ y_i f(x_i)-1=y_i(w^Tx_i+b)-1\geq 0 \\ \alpha_i(y_i f(x_i)-1)=0 \quad (\alpha_i=0 \text{ 或 } y_i f(x_i)+1=0) \end{cases}$$

解出 α 后，求出 w 与 b 即可得到模型.

$$f(x)=w^Tx+b$$
$$=\sum_{i=1}^{m}\alpha_i y_i x_i^T x+b$$



对于约束曲面上的任意点 x，该点的梯度 $\nabla g(x)$ 正交于约束曲面

对于最优点 x^*，目标函数在该点的梯度 $\nabla f(x^*)$ 正交于约束曲面.

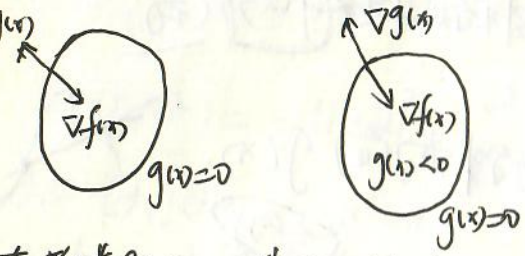
因此在最优点 $\nabla g(x)$, $\nabla f(x^*)$ 的方向必相同或相反，即存在 $\nabla f(x^*)+\lambda\nabla g(x^*)=0$

λ 为拉格朗日乘子：于是定义拉格朗日函数

$$L(x,\lambda)=f(x)+\lambda g(x)$$

其导数为 0 时的点（等高线），为极值点

KKT条件 $L(x,\lambda)=f(x)+\lambda g(x)$



在等式$g(x)=0$或在不等式约束$g(x)\le 0$下最小化目标函数$f(x)$.

$g(x)<0$:

$g(x)<0$时，对$f(x)$求极值相当于闭区间求极值，最值点即为极值点，令$\lambda=0$，直接对f求梯度即可得到极值.

$g(x)=0$:

极值点在边界取得，f的梯度与g的梯度相反，从而存在常数$\lambda>0$，使得$\nabla f(x^*)+\lambda\nabla g(x^*)=0$

$$\begin{cases} g(x)\le 0 \\ \lambda\ge 0 \\ \lambda g(x)=0 \end{cases}$$

KKT条件是目标函数在约束条件下取得极值的充要条件。目标函数在约束条件下取得极值时对应的x，λ必须满足KKT条件。

对偶问题：

由 $\min \frac{1}{2}\|w\|^2 \Rightarrow \max L(w,b,\alpha)$

对偶问题推导：

$L(x,\mu)=f(x)+\sum_k \mu_k g_k(x)$ $\mu_k\ge 0$, $g_k(x)\le 0$.

$\mu_k\ge 0$, $g_k(x)\le 0 \Rightarrow \mu_k g_k(x)\le 0$

$\max_{\mu} L(x,\mu) = f(x)$ ①

$\min_x f(x) = \min_x \max_{\mu} L(x,\mu)$ ②

$\max_{\mu}\min_x L(x,\mu) = \max_{\mu}[\min_x f(x)+\min_x \mu g(x)]$

$= \max_{\mu}\min_x f(x) + \max_{\mu}\min_x \mu g(x)$

$$\min_x \mu g(x) = \begin{cases} 0, & \mu=0 \text{ or } g(x)=0 \\ -\infty, & \mu>0 \text{ and } g(x)<0 \end{cases}$$

∵ 对 $\min \mu g(x)$ 取 max 时最大只能取0，此时 $\mu=0$ 或 $g(x)=0$ ③

∴ $\max_{\mu}\min_x \mu g(x)=0 \Rightarrow \mu=0$ or $g(x)=0$

∴ $\max_{\mu}\min_x L(x,\mu) = \min_x f(x) + 0 = \min_x f(x)$

称 $\max\min L(x,\mu)$ 为 $\min\max L(x,\mu)$ 的对偶问题，在最优 x^* 处 $\mu=0$ 或者 $= \min_x \max_{\mu} L(x,\mu)$

$$\frac{\partial L(x,\mu)}{\partial x}_{x=x^*} = 0$$

$$\begin{cases} \min_x \max_\mu L(x,\mu) = \max_\mu \min_x L(x,\mu) = \min_x f(x) = f(x^*) \\ \mu_k g_k(x) = 0 \\ \frac{\partial L(x,\mu)}{\partial x}_{x=x^*} = 0. \end{cases}$$

回到 SVM：

其 KKT：$\begin{cases} a_i \geq 0 \\ y_i f(x_i) - 1 \geq 0 \\ a_i (y_i f(x_i) - 1) = 0 \end{cases}$

当 $a_i = 0$，该样本将不会在式 $f(x) = \sum_{i=1}^{m} a_i y_i x_i^T x + b$ 的求和中出现，也不会对 $f(x)$ 有任何影响.

当 $a_i > 0$，则必有 $y_i f(x_i) = 1$

SMO (Sequential Minimal Optimization)

① 选取一对需要更新的变量 a_i 和 a_j

② 固定 a_i 和 a_j 以外的参数，求解

$$\max_a \sum_{i=1}^{m} a_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} a_i a_j y_i y_j x_i^T x_j$$

获得更新后的 a_i 与 a_j

SMO 算法之所以高效，是由于在固定其他参数后，仅优化两个参数的过程能做到非常高效

$a_i y_i + a_j y_j = C$，$a_i \geq 0$，$a_j \geq 0$

$$C = -\sum_{k \neq i,j} a_k y_k$$

使 $\sum_{i=1}^{m} a_i y_i = 0$. 用 $a_i y_i + a_j y_j = C$ 消去 a_j 可求出更新后 a_i 和 a_j.

对任意支持向量 (x_s, y_s)，都有 $y_s f(x_s) = 1$

$$y_s \left(\sum_{i \in S} a_i y_i x_i^T x_s + b \right) = 1$$

用所有支持向量求解的平均值：

$$b = \frac{1}{|S|} \sum_{s \in S} \left(\frac{1}{y_s} - \sum_{i \in S} a_i y_i x_i^T x_s \right).$$

II 核函数: Kernel function

解决非线性可分.

将样本从原始空间映射到更高维空间,使其变成线性可分问题.

如果原始空间是有限维,那么一定存在一个高维特征空间使样本可分.

令$\phi(x)$表示x映射后的特征向量,于是在特征空间中划分超平面所对应的模型可表示为:

$$f(x)=w^T\phi(x)+b$$

$$\Downarrow$$

$$\begin{cases}\min\limits_{w,b}\frac{1}{2}\|w\|^2\\ s.t.\ y_i(w^T\phi(x_i)+b)\geq 1,\quad i=1,2,\cdots,m.\end{cases}$$

$$\Downarrow$$

对偶问题:

$$\begin{cases}\max\limits_{\alpha}\sum\limits_{i=1}^{m}\alpha_i-\frac{1}{2}\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m}\alpha_i\alpha_j y_i y_j\phi(x_i)^T\phi(x_j)\\ s.t.\ \sum\limits_{i=1}^{m}\alpha_i y_i=0\quad(\alpha_i\geq 0)\end{cases}$$

$\phi(x_i)^T\phi(x_j)$是x_i与x_j映射到特征空间之后的内积

直接计算$\phi(x_i)^T\phi(x_j)$通常是困难的

$$K(x_i,x_j)=\langle\phi(x_i),\phi(x_j)\rangle=\phi(x_i)^T\phi(x_j)$$

对偶式重写为:

$$\max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i,x_j)$$

$$s.t.\ \sum_{i=1}^{m}\alpha_i y_i=0\quad(\alpha_i\geq 0,\ i=1,2,\cdots,m)$$

求解后即可得到:

$$\begin{aligned}f(x)&=w^T\phi(x)+b\\&=\sum_{i=1}^{m}\alpha_i y_i\phi(x_i)^T\phi(x)+b\\&=\sum_{i=1}^{m}\alpha_i y_i K(x,x_i)+b.\end{aligned}$$

定理: $K(\cdot,\cdot)$是定义在$\mathcal{X}\times\mathcal{X}$上的对称函数,当且仅当对于任意数据$D=\{x_1,x_2,\cdots,x_m\}$"核矩阵"$K$总是半正定:

$$K=\begin{bmatrix}K(x_1,x_1) & \cdots & K(x_1,x_n)\\ \vdots & \ddots & \vdots\\ K(x_n,x_1) & \cdots & K(x_n,x_n)\end{bmatrix}$$

由于不知道特征映射的形式，也就是不知道原本特征的样子，所以我们并不知道经过怎样的映射，也就是选择怎样的核函数，我们并不知道怎样的核函数是合适的。

III

软间隔

允许某些样本不满足约束：$y_i(w^Tx_i+b)\geq 1$

优化目标：$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} l_{0/1}(y_i(w^Tx_i+b)-1)$

当C无穷大，所有样本都必须满足 $y_i(w^Tx_i+b)\geq 1$

当C取有限值，允许一些样本不满足约束。

$$l_{0/1}(z) = \begin{cases} 1, & z<0 \\ 0, & z\geq 0 \end{cases}$$

$l_{0/1}$ 非凸，不连续，数学性质不好

hinge损失：$l_{hinge}(z) = \max(0, 1-z)$

指数损失：$l_{exp}(z) = \exp(-z)$

对率损失：$l_{log}(z) = \log(1+\exp(-z))$

采用hinge损失：$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \max(0, 1-y_i(w^Tx_i+b))$

引入"松弛变量" $\xi_i \geq 0$

$s.t.\ \xi_i \geq 1-y_i(w^Tx_i+b)$

使用拉格朗日函数：

$$f(x) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$g_1(x) = 1-y_i(w^Tx_i+b)-\xi_i \leq 0$$

$$g_2(x) = \xi_i \geq 0 \quad -\xi_i \leq 0$$

$$L(w,b,\alpha,\xi,\mu) = f(x) + \alpha_i g_1(x) + \mu_i g_2(x)$$

$$= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i(1-y_i(w^Tx_i+b)-\xi_i) - \sum_{i=1}^{m}\mu_i\xi_i$$

其中 $\alpha_i \geq 0$，$\mu_i \geq 0$ 是拉格朗日乘子

令 $L(w,b,\alpha,\xi,\mu)$ 对 w，b，ξ_i 的偏导为0

$$w = \sum_{i=1}^{m}\alpha_i y_i x_i$$

$$0 = \sum_{i=1}^{m}\alpha_i y_i$$

$$C = \alpha_i + \mu_i$$

对偶问题，代入后：

$$\begin{cases} \max\limits_{\alpha} \sum\limits_{i=1}^{m}\alpha_i - \frac{1}{2}\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j \\ \sum\limits_{i=1}^{m}\alpha_i y_i = 0 \end{cases}$$

KKT条件要求

$$\begin{cases} \alpha_i \geq 0,\ \mu_i \geq 0 \\ y_i f(x_i) - 1 + \xi_i \geq 0 \\ \alpha_i (y_i f(x_i) - 1 + \xi_i) = 0 \\ \xi_i \geq 0,\ \mu_i \xi_i = 0 \end{cases}$$

IV

核方法

学得的函数总能表示为 $f(x) = \sum_{i=1}^{m} \alpha_i y_i \kappa(x, x_i) + b$.

表示为核函数 $\kappa(x, x_i)$ 的线性组合.

任何一个核函数都隐式地定义了一个称为"再生核希尔伯特空间"

令 H 为核函数 κ 对应的再生核希尔伯特空间, $\|h\|_H$ 表示 H 空间中关于 h 的范数.

$\mapsto$: 映射

对任意单调递增函数 $\Omega: [0, \infty] \mapsto R$

任意非负损失函数 $l: R^m \mapsto [0, \infty]$

优化问题: $\min_{h \in H} F(h) = \Omega(\|h\|_H) + l(h(x_1), h(x_2), \cdots, h(x_m))$

解总可写为: $h^*(x) = \sum_{i=1}^{m} \alpha_i \kappa(x, x_i)$ ①

核线性判别分析 (KRLDA)

某种映射 $\phi: x \mapsto F$, 将样本映射到特征空间 F, 然后在 F 中执行线性判别分析, 求得:

$h(x) = w^T \phi(x)$ ②

$$\max_w J(w) = \frac{w^T S_b^\phi w}{w^T S_w^\phi w}$$

S_b^ϕ: between: 类间散度矩阵

S_w^ϕ: within: 类内散度矩阵

令 X_i 表示第 $i \in \{0, 1\}$ 类样本的集合, 样本数 m_i

第 i 类样本在特征空间 F 中的均值为:

$$\mu_i^\phi = \frac{1}{m_i} \sum_{x \in X_i} \phi(x)$$

两个散度矩阵:

$$S_b^\phi = (\mu_1^\phi - \mu_0^\phi)(\mu_1^\phi - \mu_0^\phi)^T$$

$$S_w^\phi = \sum_{i=1}^{1} \sum_{x \in X_i} (\phi(x) - \mu_i^\phi)(\phi(x) - \mu_i^\phi)^T$$

$\kappa(x, x_i) = \phi(x_i)^T \phi(x)$ ③

$w = \sum_{i=1}^{m} \alpha_i \phi(x_i)$ ①②③

令 $K \in R^{m \times m}$ 为核函数对应的核矩阵. $K_{ij} = K(x_i, x_j)$

$l_i \in \{0,1\}^{m \times 1}$, 当 $x_j \in X_i$ $l_j = 1$ (分类正确)

否则 $l_j = 0$ (分类错误)

$l_0 = (1,1,1,0,0,0)$ 属于0,为1.

$l_1 = (0,0,0,1,1,1)$ 属于1,为1.

$\hat{\mu}_0 = \frac{1}{m_0} K l_0$,

$\hat{\mu}_1 = \frac{1}{m_1} K l_1$,

$M = (\hat{\mu}_0 - \hat{\mu}_1)(\hat{\mu}_0 - \hat{\mu}_1)^T$,

$N = KK^T - \sum_{i=0}^{1} m_i \hat{\mu}_i \hat{\mu}_i^T$,

于是: $\max_w J(w) = \frac{w^T S_b^\phi w}{w^T S_w^\phi w}$

$= \max_\alpha J(\alpha)$

$= \frac{\alpha^T M \alpha}{\alpha^T N \alpha}$

2019.10.05

I 贝叶斯分类器

在样本 x 上的"条件风险"

$$R(c_i|x) = \sum_{j=1}^{N} \lambda_{ij} P(c_j|x)$$

λ_{ij}: 将 j 判为 i 的损失

$h: \mathcal{X} \mapsto \mathcal{Y}$ 最小化总体风险

$$R(h) = E_x[R(h(x)|x)]$$

$h^*(x) = \arg\min_{c \in \mathcal{Y}} R(c|x)$ 对每个样本最小化风险.

h^* 为最优贝叶斯分类器, $R(h^*)$ 为贝叶斯风险.

若最小化分类错误率,则误判损失 λ_{ij}

$$\lambda_{ij} = \begin{cases} 0, & \text{若 } i=j \\ 1, & \text{otherwise} \end{cases}$$

条件风险: $R(c|x) = 1 - P(c|x)$

$$h^*(x) = \arg\max_{c \in \mathcal{Y}} P(c|x)$$

对每个样本 x,使后验概率 $P(c|x)$ 最大

生成模型必然考虑：

$$P(c|x)=\frac{P(x,c)}{P(x)}$$

$$P(c|x)=\frac{P(c)\,P(x|c)}{P(x)}$$

$P(c)$：样本空间中各类样本所占的比例

极大似然估计：源自频率主义学派，根据数据条样来估计概率分布参数.

$$P(D_c|\theta_c)=\prod_{x\in D_c}P(x|\theta_c)$$ （使用θ_c找到D_c中所有样本的概率）

在所有θ_c的可能取值中，找到一个使$P(D_c|\theta_c)$最大的值.

$$LL(\theta_c)=\log P(D_c|\theta_c)=\sum_{x\in D_c}\log P(x|\theta_c)$$

$$p(x|c)\sim N(\mu_c,\sigma_c^2)$$

11

朴素贝叶斯分类器

采用"属性条件独立性假设"：对已知类别，假设所有属性相互独立，每个属性独立地对分类结果发生影响.

$$P(c|x)=\frac{P(c)\,P(x|c)}{P(x)}$$

x=青绿，大，西瓜

x_1=青绿

x_2=大

x_3=西瓜

$$=\frac{P(c)}{P(x)}\prod_{i=1}^{d}P(x_i|c)$$

$$P(x|c)=P(x_1|c)\,P(x_2|c)\,P(x_3|c)$$

$$=P(x_1,x_2,x_3|c)$$

朴素贝叶斯判定准则：

$$h_{nb}(x)=\arg\max_{c\in y}P(c)\prod_{i=1}^{d}P(x_i|c)$$

基于训练集D，估计类先验概率$P(c)$

$$P(c)=\frac{|D_c|}{|D|}$$

D_{c,x_i}表示D_c中在第i个属性上取值为x_i的样本组成的集合.

$$P(x_i|c)=\frac{|D_{c,x_i}|}{}$$

为了避免其他属性携带的信息被训练集中未出现的属性值"抹去"，在估计概率值时通常要进行"平滑"，常用"拉普拉斯修正"

N表示训练集D中可能的类别数.

$$P(c)=\frac{|D_c|}{|D|}$$

$$\hat{P}(c)=\frac{|D_c|+1}{|D|+N}$$

$$P(x_i|c)=\frac{|D_{c,x_i}|}{|D_c|} \Rightarrow \hat{P}(x_i|c)=\frac{|D_{c,x_i}|+1}{|D_c|+N_i}$$

半朴素贝叶斯分类器.

对"属性条件独立性假设"进行一定程度的放松

III.

EM算法

Z表示隐变量集

Θ表示模型参数

X表示已观测变量.

$$LL(\Theta|X,Z)=\ln P(X,Z|\Theta)$$

$$LL(\Theta|X)=\ln P(X|\Theta)=\ln \sum_Z P(X,Z|\Theta)$$

EM常用来估计参数隐变量，迭代式的方法.

以初始值Θ^0为起点：

E① 基于Θ^t推断隐变量Z的期望，记为Z^t

M② 基于已观测变量X和Z^t对参数Θ做极大似然估计，记为Θ^{t+1}

直到 Θ^t和Z^t稳定

基于Θ^t计算隐变量Z的概率分布$P(Z|X,\Theta^t)$

E① 以当前Θ^t推断$P(Z|X,\Theta^t)$，并计算对数似然

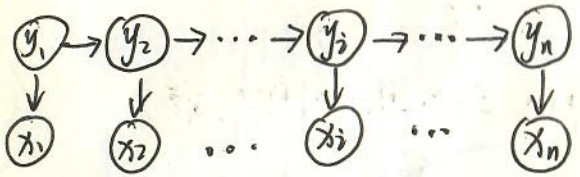
$$LL(\Theta|X,Z)$$

$$Q(\Theta|\Theta^t)=E_{Z|X,\Theta^t}LL(\Theta|X,Z)$$

2019.10.07.

概率图模型.

隐马尔可夫模型的图结构



所有变量的联合概率分布为:

$$P(x_1, y_1, \cdots, x_n, y_n) = P(y_1)P(x_1|y_1)\prod_{i=2}^{n} P(y_i|y_{i-1})P(x_i|y_i)$$

确定一个隐马尔可夫模型还需:

① 状态转移概率: $A=[a_{ij}]_{N\times N}$

$a_{ij} = P(y_{t+1}=s_j \mid y_t=s_i)$, $1\le i,j\le N$.

② 输出观测概率: $B=[b_{ij}]_{N\times M}$

$b_{ij} = P(x_t=o_j \mid y_t=s_i)$, $1\le i\le N$, $1\le j\le M$

$M \le N$, 因为有些状态并不能转换为任意其他状态.

③ 初始状态概率: $\pi=(\pi_1, \pi_2, \pi_3, \cdots, \pi_N)$

$\pi_i = P(y_1=s_i)$, $1\le i\le N$.

HMM的问题

① 给定模型 $\lambda=[A, B, \pi]$, 如果计算 $x=\{x_1, x_2, \cdots, x_n\}$ 的概率 $P(x|\lambda)$

② 给定模型 $\lambda=[A, B, \pi]$ 和观测序列 $x=\{x_1, x_2, \cdots, x_n\}$, 找出与此观测序列最匹配的状态序列 $y=\{y_1, y_2, \cdots y_n\}$

③ 给定观测序列 $x=\{x_1, x_2, \cdots, x_n\}$, 调整模型参数 $\lambda=[A, B, \pi]$, 使 $P(x|\lambda)$ 最大.

马尔可夫随机场

n个变量 $x=\{x_1, x_2, \cdots, x_n\}$

所有团构成的集合为 C

与团 $Q\in C$ 对应的变量集合记为 x_Q

联合概率 $P(x)$:

$$P(x) = \frac{1}{Z}\prod_{Q\in C}\psi_Q(x_Q)$$

ψ_Q 为团 Q 对应的势函数.

$Z = \sum_x \prod_{Q\in C}\psi_Q(x_Q)$

若团 Q 等于 Q^* 极大团.

$P(X_A, X_B | X_C) = P(X_A | X_C) P(X_B | X_C)$

X_A 和 X_B 在给定 X_C 时条件独立.

局部马尔可夫性: $X_v \perp X_{V \setminus n^*(v)} \mid X_{n(v)}$

给定结点 v 的邻近结点 $X_{n(v)}$, 则结点 v 与除结点 v 及 $X_{n(v)}$ 之外的所有结点 $X_{V \setminus n^*(v)}$ 相互独立.

成对马尔可夫性: 给定所有其他变量, 两个非邻接变量条件独立.

条件随机场. ~19. 10. 14.

判别式模型

MCMC: (Markov Chain Monte Carlo)

计算 $f(x)$ 在 $p(x)$ 上的期望

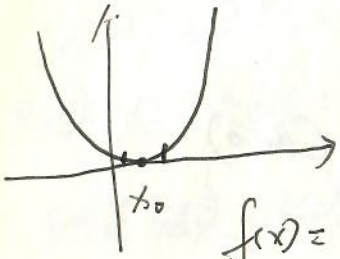
$$P(A) = \int_A p(x)\,dx$$

$$P(f) = E_p[f(x)] = \int_x f(x) p(x)\,dx$$

$$\hat{P}(f) = \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

泰勒展开式 (3Blue1Brown)

函数在某点的各阶导数与构造的多项式的对应导数相同.



$y = f(x)$

$y_0 = f(x_0)$

$f(x) = \underset{\Delta x \to 0}{f(x_0 \pm \Delta x)} = y_0 \pm f'(x_0 \pm \Delta x)$

由于 $y_0 = f(x_0)(x-x_0)^0$

$f(x) = f^{(1)}(x_0)(x-x_0)^0 + f'(x_0)(x-x_0)^1$

题目: $f(x) = \sum_{i=0}^{n} f^{(i)}(x_0)(x-x_0)^2$

$f'(x) = f'(x_0)$

$f''(x) = 2f''(x_0)$??

对 $f^{(m)}(x_0)(x-x_0)^m$ 求导看看:

m阶导: $m! f^{(m)}(x_0)$ m+1阶导: 0.

一m阶导一: m阶导以后都是0, m阶导之前 $(x-x_0)$ 都是0.

对 $f(x)$ 求 m 阶导为 $m! f^{(m)}(x)$

2019.11.15.

为什么梯度下降更新参数有效.

$$\Delta w_{hj} = -\eta \frac{\partial E_k}{\partial w_{hj}}$$

$$w_{hj} = w_{hj} + \Delta w_{hj}$$

$$= w_{hj} - \eta \frac{\partial E_k}{\partial w_{hj}}$$

η: 学习率

$\frac{\partial E_k}{\partial w_{hj}}$: 梯度.

如果一个函数n阶可导,那么我们可以用多项式仿造一个相似的函数,这就是泰勒展开式.

$$f(x)_{Taylor} = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} \times (x-a)^n$$

$$= f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f^{(2)}(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_n(x)$$

$$f(x)_{Taylor} = \sum_{i=0}^{n} \frac{f^{(i)}(x_0)}{i!}(x-x_0)^i, \quad x \in (x_0 - \Delta x, x_0 + \Delta x)$$

无限逼近

用一阶泰勒展开式证明梯度下降：

$f(\theta)=f(\theta_0)+f'(\theta_0)\cdot(\theta-\theta_0)$

设θ_0是原点，θ是向下步后所处的点。

min：$f(\theta)-f(\theta_0)=(\theta-\theta_0)\cdot f'(\theta_0)$

用向量$\vec{v}$表示$\theta-\theta_0$：$\vec{v}=\theta-\theta_0$

$f(\theta)-f(\theta_0)=\vec{v}\cdot f'(\theta_0)=\|\vec{v}\|\cdot\|f'(\theta_0)\|\cdot\cos\alpha$

要使$f(\theta)-f(\theta_0)$最小，$\cos\alpha=-1$，$\vec{v}$与$f'(\theta_0)$方向相反。

$\vec{v}$：$\vec{v}=-\eta\cdot f'(\theta_0)$

因为：$\vec{v}=\theta-\theta_0$，$\theta-\theta_0=-\eta\cdot f'(\theta_0)$

$$\theta=\theta_0-\eta\cdot f'(\theta_0)$$

因为要满足一阶泰勒展开式，所以

$\theta\to\theta_0$，$\eta\cdot f'(\theta_0)$要小，η要小，学习率要小。

生信中的多元统计

矩阵·$\begin{bmatrix} m \\ n \end{bmatrix}$：对 $\begin{bmatrix} m \\ n \end{bmatrix}$ 的线性变换.

$\begin{bmatrix} m \\ n \end{bmatrix} = m\begin{bmatrix} 1 \\ 0 \end{bmatrix} + n\begin{bmatrix} 0 \\ 1 \end{bmatrix}$

矩阵·矩阵：对多个变量(样本)同时做相同的线性变换.
(多变量组合)

矩阵的行列式 $\det(A)$ 表示矩阵的面积，体积等.

正交矩阵：方阵 $AA^T = I$

正交向量：$a^T b = a_1b_1 + a_2b_2 + \cdots + a_pb_p = 0$

如果某方阵 A 为正交方阵

$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$ 向量两两正交，且 $\|X_1\| = \sqrt{a_{11}^2 + a_{12}^2 + a_{13}^2}$ 为单位向量

① 相似矩阵：

$A = PBP^{-1}$，A 与 B 互为相似矩阵，有相同特征值.

② 若 A(n×n) 有 n 个线性无关的特征向量，那么 A 矩阵可以被对角化 $|A| \neq 0$

$A = P\Lambda P^{-1}$，Λ 为对角矩阵，$\begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_n \end{bmatrix}$ 特征值矩阵.

③ 若 A 为对称矩阵，有 n 个线性无关特征向量. $|A| \neq 0$

$A = T\Lambda T^{-1} = T\Lambda T^T$ Λ 为特征值矩阵

T 正交矩阵，特征向量矩阵.

$= (t_1, t_2, \cdots, t_p)\begin{pmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_p \end{pmatrix}\begin{pmatrix} t_1' \\ t_2' \\ t_p' \end{pmatrix} = \sum_{i=1}^{p} \lambda_i t_i t_i^T$

(特征值为解析)

概率论与数理统计

pdf(f(x)) cdf = F(x) $\int_{-\infty}^{x} f(x)dx = F(x)$

$F(x) = P(X \leq x)$ $X \sim F(x)$

离散型，$F(x) = \sum_{x_i \leq x} p_i$

连续型，$F(x) = \int_{-\infty}^{x} f(x)dx$.

pdf：概率分布函数 $f(x) = F'(x)$

cdf：累积分布函数 $F(x)$

① 均匀分布 $f(x) = C$ $F(x) = Cx$

② 二项分布 $C_n^k p^k (1-p)^{n-k}$ $\sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k}$ $X \sim B(n,p)$

③ 负二项分布，连续不断且独立地重复进行一个参数为 p 的伯努利试验，记 X 为第 r 次"成功"所需试验次数. k 次试验中，前 k-1 次成功 r-1 次.

$P(X=k) = p \cdot P(B(k-1, p) = r-1)$

$= pC_{k-1}^{r-1} p^{r-1} q^{k-1-(r-1)} = C_{k-1}^{r-1} p^r q^{k-r}$.

$X \sim NB(n, p)$，限制了第 k 次为成功的 r 次.

$q = 1-p$，$k = r, r+1, \cdots$

④ 泊松分布 $P_k = P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}$，$k = 0, 1, 2, \cdots$

$0! = 1$ $\sum_{k=0}^{\infty} P_k = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1$

泊松分布与二项分布的关系：

$X \sim B(n,p)$，$Y \sim P(np)$ $\lambda = np$

p很小，n很大。

$$\frac{B_k(n,p)}{B_{k-1}(n,p)} = \frac{C_n^k p^k (1-p)^{n-k}}{C_n^{k-1} p^{k-1} (1-p)^{n-k+1}} = \frac{\frac{n!}{(n-k)!k!} p^k (1-p)^{n-k}}{\frac{n!}{(n-k+1)!(k-1)!} p^{k-1} (1-p)^{n-k+1}} = \frac{(n-k+1)p}{k(1-p)}$$

$$= \frac{np - kp + p}{k(1-p)} = \frac{\lambda - (k-1)p}{k(1-p)} \approx \frac{\lambda}{k}$$

$$B_0(n,p) = (1-p)^n = (1-\frac{\lambda}{n})^n = (1+(-\frac{\lambda}{n}))^{(-\frac{n}{\lambda})\cdot(-\lambda)} \approx e^{-\lambda}$$

（泰勒展开式）

几何分布：（无记忆离散）

独立重复参数为P的伯努利分布，X为首次成功。

$P(X=k) = (1-p)^{k-1} p \quad 0 < p < 1$

$X \sim Ge(p)$

指数分布：（无记忆连续）

$X \sim Exp(\lambda)$，$\lambda > 0$ $F(x) = P(x \le x)$

$$f(x) = \begin{cases} 0 & x \le 0 \\ \lambda e^{-\lambda x} & x > 0 \end{cases} \quad = \int_{-\infty}^{x} \lambda e^{-\lambda x} dx \quad \lambda x$$

正态分布：$X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma}} \quad (pdf)$$

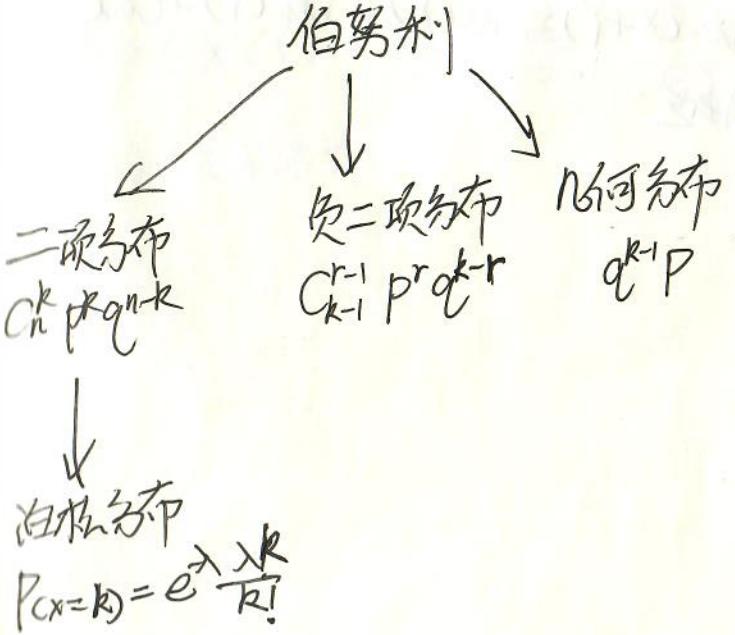
$\mu = 0$，$\sigma^2 = 1$，$X \sim N(0,1)$ 标准正态分布。

若 $X \sim N(\mu, \sigma^2)$，$\frac{X-\mu}{\sigma^2} \sim N(0,1)$

数据分析时的标准化可能就是转换为标准正态分布。

卡方分布：

$X \sim N(0,1)$，X^2 称为一个自由度的 χ^2 分布。



随机变量X函数期望.

$$E(g(X))=\begin{cases}\sum_{k=1}^{\infty} g(x_k)\cdot P(X=x_k) & \text{离散}\\ \int_{-\infty}^{+\infty} g(x)f(x)dx & \text{连续}\end{cases}$$

$Var(X)$ 方差：随机变量偏离期望的程度.

$$\begin{aligned}Var(X)&=E((X-E(X))^2)\\&=E(X^2-2XE(X)+E(X)^2)\\&=E(X^2)-2E(X)E(X)+E(X)^2\\&=E(X^2)-E(X)^2\end{aligned}$$

X与Y独立：$Var(X+Y)=Var(X)+Var(Y)$

X与Y不独立：$Var(X+Y)=Var(X)+Var(Y)+Cov(X,Y)$

不相关可以不独立

协方差

$$\begin{aligned}Cov(X,Y)&=E[(X-E(X))(Y-E(Y))]\\&=E[XY-XE(Y)-YE(X)+E(X)E(Y)]\\&=E(XY)-E(X)E(Y)-E(Y)E(X)+E(X)E(Y)\\&=E(XY)-E(X)E(Y)\end{aligned}$$

$$corr(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)}\cdot\sqrt{Var(Y)}}=\frac{Cov(X,Y)}{\sigma_X\cdot\sigma_Y}=R.$$

$$Var\left(\frac{X}{\sigma_X}\right)=1\quad Var\left(\frac{Y}{\sigma_Y}\right)=1$$

$$corr(X,Y)=Cov\left(\frac{X}{\sigma_X},\frac{Y}{\sigma_Y}\right)$$

马尔科夫不等式：

假设检验：

第一类错误：α：拒绝真

第二类错误β：接受假.

原假设H_0一般用不好的结果，备择假设H_1用好的

PCA理解（通俗解）

对A矩阵做PCA，线性转换.

$PA=C$ 转换矩阵为P（P为正交矩阵，因为投影方向垂直）

C的协方差矩阵

$D=Cov(C,C)=\frac{CC^T}{n-1}=Var(C)$

同方向有协方差，不同方向协方差为0.

$D=Var(C)=Var(PA)=PVar(A)P^T$

$=PAA^TP^T$

$=PBP^T \quad B=AA^T$

$D=PBP^T=PBP^{-1}$

$P^{-1}DP=B$ D（特征值矩阵）

P（特征向量矩阵）

（这里根据定义P正交，D对角，则P为B的特征向量矩阵，D为B的特征值矩阵）.

R语言：

	gene1	gene2	P
sample1			
sample2			
n			

$\rightarrow A^T_{nxp}=A^T_{pxn}$

A_{pxn}

$A^TA=B$

PCA之前可以挑选高变异的gene.
挑选方差，CV高的基因.
500, 1000, 3000个基因.

PCA的关键步骤及理解
① 获得数据，矩阵A进行归一化
② AA^T为A矩阵的协方差矩阵，记作B.
③ 对A正交旋转不影响A的协方差总和
④ B正交对角化 $B=QDQ^{-1}=QDQ^T$ $Q^TQ=QQ^T=I$

聚类
层次聚类：
hierarchical clustering method.
聚集法：按距离，每次设置cutoff，最后聚成一类.
分割法：由少到多

① cutoff设置？
② 计算距离？
③ 选择参考点

明考夫斯基距离

$$d(x,y)=\left[\sum_{i=1}^{p}|x_i-y_i|^p\right]^{1/p}$$

当$q=1$时，为绝对值距离（曼哈顿距离，街区距离）

$$d(x,y)=\sum_{i=1}^{p}|x_i-y_i|$$

当$q=2$时，为欧氏距离

当$q=\infty$时，为切比雪夫距离

$$d(x,y)=\max_{1\le i\le p}|x_i-y_i|$$

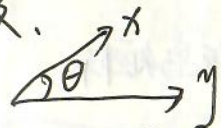
兰氏距离（Lance and williams distance）

所有数据为正，$d(x,y)=\sum_{i=1}^{p}\frac{|x_i-y_i|}{x_i+y_i}$

该距离与各变量的单位无关，且适用于高度偏斜或含异常值的数据.

定义 x, y 之间的距离满足：
① 非负性：$d(x,y)\ge 0$
② 对称性：$d(x,y)=d(y,x)$
③ 三角不等式：$d(x,y)\le d(x,z)+d(z,y)$

相似系数.

夹角余弦 

$\cos\theta=\frac{x^Ty}{\|x\|\|y\|}$

Pearson相关系数 $\rho=\rho(x,y)=\frac{Cov(x,y)}{\sqrt{V(x)\cdot V(y)}}$

相似系数与距离之间的关系

距离 ⇒ 相似系数

$$c_{ij} = \frac{1}{1+d_{ij}}$$

相似系数 ⇒ 距离　相似系数矩阵 (c_{ij}) 非负定.

$$d_{ij} = \sqrt{2(1-c_{ij})}$$

dist (距离矩阵) → hclust (层次聚类)

(WGCNA也是层次聚类, 聚 gene module).

MDS (多维标度法)

找出因子的矩阵 D (距离), 在低维空间中寻找一个矩阵 X, 使得 X 构成的欧式距离矩阵 B 尽可能和原来矩阵保持一致. 此种多维标度法也称作主座标分析.

① 样本矩阵 $X = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} = [\vec{x}_1, \vec{x}_2, \cdots \vec{x}_n]$

m个sample, n个gene.

② $D = \begin{bmatrix} d(x_1,x_1) & \cdots & d(x_1,x_m) \\ \vdots & & \\ d(x_m,x_1) & & d(x_m,x_m) \end{bmatrix}$ 距离矩阵.

③ $\vec{x}_i \to \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$, $\vec{y}_i \begin{bmatrix} y_{i1} \\ \vdots \\ y_{ip} \end{bmatrix}$? $\vec{y}_i$ 产生的距离矩阵 B

④ 找到 $Y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{ip} \end{bmatrix}$, 令 $B = Y^T Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{bmatrix} [Y_1, Y_2, \cdots Y_m]$

中心化　(B是为了方便计算)

⑤ $dist_{ij} = d(x_i, x_j) = \|\vec{Y}_i - \vec{Y}_j\|$ (欧氏距离)

$$dist_{ij}^2 = \|Y_i\|^2 + \|Y_j\|^2 - 2\vec{Y}_i \cdot \vec{Y}_j$$

$$dist_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij}$$

$$b_{ij} = \frac{1}{2}(b_{ii} + b_{jj} - dist_{ij})$$

⑥ $\sum_{i=1}^m b_{ij} = \sum_{i=1}^m \vec{Y}_i \cdot \vec{Y}_j = \sum_{i=1}^m = \sum_{i=1}^m (y_{i1}y_{1j} + y_{i2}y_{2j} + \cdots + y_{ip}y_{pj})$

$$= y_{11}y_{1j} + y_{21}y_{2j} + \cdots + y_{p1}y_{pj} +$$
$$y_{12}y_{1j} + y_{22}y_{2j} + \cdots + y_{p2}y_{pj} +$$
$$\vdots \qquad \vdots \qquad \vdots$$
$$y_{1m}y_{1j} + y_{2m}y_{2j} + \cdots + y_{pm}y_{pj}$$

$$= y_{1j} \underbrace{\sum_{i=1}^m y_{i2}}_{0} + y_{2j} \underbrace{\sum_{i=1}^m y_{i2}}_{0} + \cdots + y_{pj} \underbrace{\sum_{i=1}^m y_{pi}}_{0}$$

$$= 0$$

⑦ $\sum_{i=1}^m dist_{ij}^2 = \sum_{i=1}^m (b_{ii} + b_{jj} - 2b_{ij}) = \sum_{i=1}^m b_{ii} + \sum_{i=1}^m b_{jj} - 2\sum_{i=1}^m b_{ij}$

$$= tr(B) + m b_{jj} + 0$$

$$\sum_{j=1}^m dist_{ij}^2 = tr(B) + m b_{ii}$$

$$\sum_{i=1}^m \sum_{j=1}^m dist_{ij}^2 = \sum_{j=1}^m (tr(B) + m b_{jj}) = \sum_{j=1}^m tr(B) + m \sum_{j=1}^m b_{jj}$$

$tr(B) = \frac{1}{2m}\sum_{i=1}^{m}\sum_{j=1}^{m} dist_{ij}^2$

$\sum_{j=1}^{m} dist_{ij}^2 = tr(B) + m b_{jj}$

$\sum_{i=1}^{m} dist_{ij}^2 = tr(B) + m b_{ii}$

$b_{ij} = -\frac{1}{2}(b_{ii} + b_{jj} - dist_{ij}^2)$

又由于：$B = Y^T Y$，B是对称矩阵，正交对角化

$B = Q \Lambda Q^T = Y^T Y$

$Y^T = Q \Lambda^{\frac{1}{2}}$

主要思想推导：

$dist_{ij} = d(x_i, x_j) = \|Y_i - Y_j\|$

$B = Y^T Y$，$B_{ij} = b_{ij} = d(Y_i, Y_j) = \|Y_i - Y_j\|$ ✗

（方便计算）.

注意 ①当 dist 矩阵本身为欧氏距离 MDS = PCA

②MDS 强调对不同种类 dist 都可转为欧氏距离.

t-SNE.

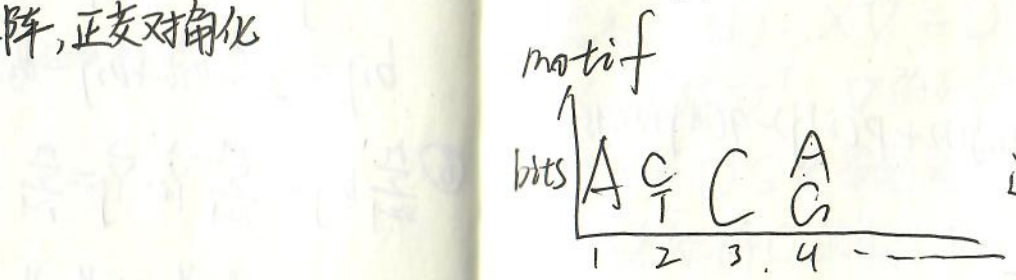
①高维空间概率分布 P_1

②低维空间概率分布 P_2

③使 P_1 与 P_2 尽可能相同，通过迭代.

香农信息熵

$H = -\sum_{i=1}^{n} P_i \log(P_i)$



通过信息熵计算.

KL散度：衡量两个分布是否相似

$D_{KL}(P\|Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}$

当两个概率分布完全相同时，$D_{KL} = 0$，否则 $D_{KL} > 0$.

SNE 的数学过程.

1. 先通过欧氏距离表示点的概率分布.

高维情况：$P_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2/(2\sigma_i^2))}$（以 x_j 为中心 / 以非 x_j 为中心）

低维情况：$q_{ij} = \frac{\exp(-\|y_i - y_j\|^2)}{}$

2. 希望高维和低维之间的分布尽可能相同.

$$C=\sum_i KL(P_i\|Q_i)=\sum_i\sum_j p_{j|i}\log\frac{p_{j|i}}{q_{j|i}}$$ (最小化)

3. 初始化时需要设置参数困惑度(perplexity), SNE会通过二分搜索的方法计算最佳的初始方差.
一般困惑度选择5-50.

4. 对目标函数求梯度

$$\frac{\partial C}{\partial y_i}=2\sum_j(p_{j|i}-q_{j|i}+p_{i|j}-q_{i|j})(y_i-y_j)$$

5. 迭代

$$Y^{(t)}=Y^{(t-1)}+\eta\frac{\partial C}{\partial Y}+\alpha(t)(Y^{(t-1)}-Y^{(t-2)})$$

t-SNE

① $p_{ij}=\frac{p_{j|i}+p_{i|j}}{2n}$

② $q_{ij}=\frac{(1+\|y_i-y_j\|^2)^{-1}}{\sum_{k\neq l}(1+\|y_k-y_l\|^2)^{-1}}$ (t分布) (近的越近,远的越远)

受初值影响. (RtSNE).

PCA理解 (from Cross Validated)

$X(n\times p)$

$C=X^TX$ (协方差矩阵应该是 $Cov=\frac{X^TX}{n-1}$)

$C=VLV^T$ (对角化,对称矩阵正交对角化)

$X=USV^T$ (奇异值分解)

$C=X^TX=VSU^TUSV^T=VS^2V^T=VLV^T$

$\lambda_i=S_i^2$

主成分(坐标): $XV=USV^TV=US$.

① $X=USV^T$, V的列向量是特征向量,是投影方向.

② XV或US是投影后的坐标

③ 协方差矩阵C的特征值 $\lambda_i=S_i^2$

④ 标准化的(除以特征值的)分数是U的列向量,从US为主成分坐标可看出.

⑤ VS是loading(载荷,变量与主成分的相关系数)

⑥ 以上的条件是行为样本,列为变量. 否则U,V变换.

⑦ PCA on correlation matrix, X的列需要中心化+标准化.

⑧ $X_K=U_KS_KV_K^T$ (U前K列, S前$K\times K$, V^T前K行).
X_K为包含前K个PCs(主成分)的压缩数据.

⑨ 若 $n>P$, $U_{n\times n}$, $V_{P\times P}$, U的最后$n-P$列随意,对应的S的行为0.

多元线性回归

偏最小二乘回归 ☆
主成分回归
} 变量多时效果较好.

一元线性回归 $R^2 = r^2$

多元线性回归最小二乘: $R^2 = \frac{SSR}{SST}$

$b = (X^TX)^{-1}X^Ty$，如果X有共线向量 b 无解.

如果X中有变量强相关，b估计不稳.

判断是否存在共线性:

tolerance $= 1 - R_j^2$ (第j个变量).

VIF (Variance inflation factor)

$$VIF_j = \frac{1}{1 - R_j^2}$$

Kappa: $K = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$ λ_i为 X^TX 的特征值.

(kappa > 15) 或 VIF ≥ 5 有强共线性

三，线性回归

$f(w) = w^T x$ （当数据满足正态分布，最小二乘就是最大似然）

噪声

Loss function: $L(w) = \sum_{i=1}^{N} \|w^T x_i - y_i\|^2$

$\frac{dL(w)}{dw} \Rightarrow \hat{w} = (X^T X)^{-1} X^T Y$

$X_{N \times p}$，N个样本，$x_i \in \mathbb{R}^p$ $N \gg p$

过拟合 ⇒ ① 加数据
② 特征选择/特征提取 (PCA)
③ 正则化

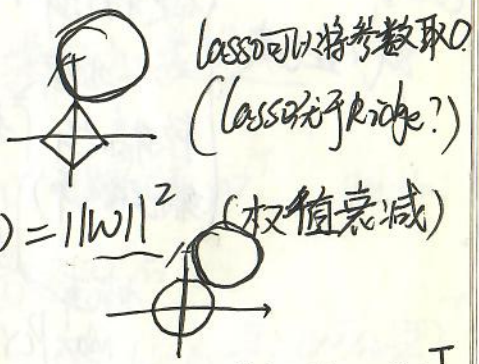
正则化框架

$\text{argmin}[L(w) + \lambda P(w)]$

loss　penalty.

L_1: Lasso, $P(w) = \|w\|$ （lasso可以将参数取0）

L_2: Ridge, 岭回归 $P(w) = \|w\|^2$ （权值衰减）



$X = (x_1, x_2 \cdots x_n)^T$

推导岭回归：

$\text{argmin}_w \sum_{i=1}^{N} \|w^T x_i - y_i\|^2 + \lambda w^T w$

$(w^T x_1 - y_1, w^T x_2 - y_2 \cdots w^T x_n - y_n) \begin{pmatrix} w^T x_1 - y_1 \\ w^T x_2 - y_2 \\ \vdots \\ w^T x_n - y_n \end{pmatrix} = w^T (x_1, x_2 \cdots x_n) - (y_1, y_2 \cdots y_n)$

$= (w^T X^T - Y^T)(Xw - Y) + \lambda w^T w = w^T X^T X w - 2 w^T X^T Y + \lambda w^T w + Y^T Y$

$\hat{w} = \arg\min J(w)$

$\frac{\partial J(w)}{\partial w} = 2(X^TX + \lambda I)w - 2X^TY = 0$

$\hat{w} = (X^TX + \lambda I)^{-1}X^TY$

λI：惩罚项.

最小二乘估计 LSE $\Longleftrightarrow$ 极大似然估计 MLE (noise 为 Gaussian 分布)

Regularized LSE $\Leftrightarrow$ MAP (最大后验) (noise 和 prior 也是 GD)

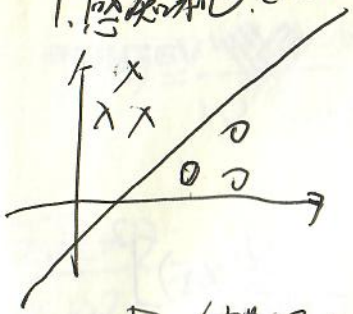
(四) 线性分类

- 硬输出：{线性判别 / 感知机} (输出结果)
- 软输出 (输出概率)
  - 判别：Logistic regression $P(Y|X)$
  - 生成 $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$

概率 $\max P(Y|X) \Rightarrow \max\left(P(X|Y)P(Y)\right)$

后验 ∝ 似然 × 先验.

1. 感知机 (Perceptron Learning Algorithm).



思想：错误驱动

模型：$f(x) = \text{sign}(w^Tx)$, $x \in \mathbb{R}^p$, $w \in \mathbb{R}^p$

Loss function:

$L(w) = \sum_{i=1}^{N} I\{y_i w^T x_i < 0\}$

D：{错误分类的样本}

$L(w) = \sum_{x_i \in D} -y_i w^T x_i$

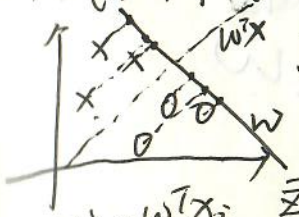
$\frac{\partial L(w)}{\partial w} = \sum_{x_i \in D} -y_i x_i$

SGD: $w^{(t+1)} = w^{(t)} - \lambda \frac{\partial L(w)}{\partial w}$

$= w^{(t)} + \lambda y_i x_i$ (改变斜率).

(LDA)

2. 线性判别分析. 思想：类内小，类间大.



threshold

投影方向 w，分离性 w^T $\|w\|=1$

$z_i = w^Tx_i$ $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i = \frac{1}{N}\sum_{i=1}^{N} w^Tx_i$

$S_z = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(z_i - \bar{z})^T = \frac{1}{N}\sum_{i=1}^{N}(w^Tx_i - \bar{z})(w^Tx_i - \bar{z})^T$

$C_1: \bar{z}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} w^Tx_i$ $S_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}(w^Tx_i - \bar{z}_1)(w^Tx_i - \bar{z}_1)^T$

$C_2: \bar{z}_2 = \frac{1}{N_2}\sum_{i=1}^{N_2} w^Tx_i$ $S_2 = \frac{1}{N_2}\sum_{i=1}^{N_2}(w^Tx_i - \bar{z}_2)(w^Tx_i - \bar{z}_2)^T$

类间：$(\bar{z}_1 - \bar{z}_2)^2$ ↑大

目的是：类内小，类间大.

目标函数：$J(w)=\frac{(\bar{z}_1-\bar{z}_2)^2}{s_1+s_2}$

$\hat{w}=\arg\max_w J(w)$

$(\bar{z}_1-\bar{z}_2)^2=\left[w^T\left(\frac{1}{N_1}\sum_{i=1}^{N_1}x_i-\frac{1}{N_2}\sum_{i=1}^{N_2}x_i\right)\right]^2$

$=[w^T(\bar{X}_{c1}-\bar{X}_{c2})]^2$

$s_1+s_2=\frac{1}{N_1}\sum_{i=1}^{N_1}(w^Tx_i-\bar{z}_1)(w^Tx_i-\bar{z}_1)^T+\frac{1}{N_2}\sum_{i=1}^{N_2}(w^Tx_i-\bar{z}_2)(w^Tx_i-\bar{z}_2)^T$

$=\frac{1}{N_1}\sum_{i=1}^{N_1}(w^Tx_i-\frac{1}{N_1}\sum_{j=1}^{N_1}w^Tx_j)(w^Tx_i-\frac{1}{N_1}\sum_{j=1}^{N_1}w^Tx_j)^T+\cdots$

$=\frac{1}{N_1}\sum_{i=1}^{N_1}w^T(x_i-\bar{X}_{c1})(x_i-\bar{X}_{c1})^Tw$

$=w^T\frac{1}{N_1}\sum_{i=1}^{N_1}(x_i-\bar{X}_{c1})(x_i-\bar{X}_{c1})^Tw$

$=w^TS_{c1}w+w^TS_{c2}w.$

$J(w)=\frac{[w^T(\bar{X}_{c1}-\bar{X}_{c2})]^2}{w^TS_{c1}w+w^TS_{c2}w}$

$=\frac{w^T(\bar{X}_{c1}-\bar{X}_{c2})(\bar{X}_{c1}-\bar{X}_{c2})^Tw}{w^T(S_{c1}+S_{c2})w}$

S_b：between-class 类间方差 $(\bar{X}_{c1}-\bar{X}_{c2})(\bar{X}_{c1}-\bar{X}_{c2})^T$

$J(w)=\frac{w^TS_bw}{w^TS_ww}=w^TS_bw(w^TS_ww)^{-1}$

$\frac{\partial J(w)}{\partial w}=2S_bw(w^TS_ww)^{-1}+w^TS_bw\cdot(-1)\cdot(w^TS_ww)^{-2}\cdot 2w$

$=0$

$S_bw(w^TS_ww)^{-1}=w^TS_bw(w^TS_ww)^{-2}\cdot w$

$S_bw(w^TS_ww)=\underbrace{w^TS_bw}_{\in\mathbb{R}}S_ww$

$w=p\times 1$

$w^T=1\times p$

$S_w=p\times p$

$\hat{w}=\frac{w^TS_ww}{w^TS_bw}S_w^{-1}\cdot S_b\cdot w$

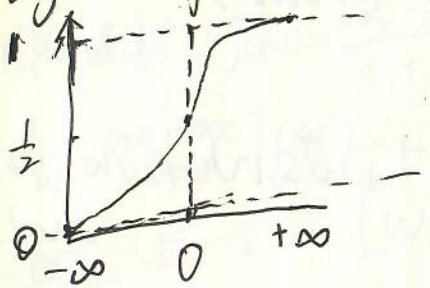
(正比于) $\propto S_w^{-1}\underbrace{S_b\cdot w}$

$(\bar{X}_{c1}-\bar{X}_{c2})\underbrace{(\bar{X}_{c1}-\bar{X}_{c2})^T\cdot w}_{\in\mathbb{R}}$

$\propto S_w^{-1}(\bar{X}_{c1}-\bar{X}_{c2})$

Logistic Regression. (广义线性模型)

sigmoid function: $\sigma(z) = \frac{1}{1+e^{-z}}$



$$P_1 = P(y=1|x) = \sigma(w^T x) = \frac{1}{1+e^{-w^T x}}, \quad y=1$$

$$P_0 = P(y=0|x) = 1 - P(y=1|x) = \frac{e^{-w^T x}}{1+e^{w^T x}}, \quad y=0$$

$$P(y|x) = P_1^{y} P_0^{1-y}$$

MLE: $\hat{w} = \arg\max_w \log P(Y|X)$

Maximum Likelihood Estimate. 最大似然估计.

$$= \arg\max_w \log \prod_{i=1}^{N} P(y_i|x_i)$$

$$= \arg\max_w \sum_{i=1}^{N} \log P(y_i|x_i)$$

$$= \arg\max_w \sum_{i=1}^{N} \overbrace{y_i \log P_1}^{\log} + \overbrace{(1-y_i) P_0}^{\log}$$

$$= \arg\max_w \sum_{i=1}^{N} y_i \log P_1 + (1-y_i) \log P_0$$

MLE $\max \Rightarrow$ loss function $\min$.

概率生成模型.

Gaussian Discriminant Analysis.

$$P(y|x) \propto P(x|y) P(y)$$

$$\hat{y} = \arg\max_{y \in \{0,1\}} P(y|x) = \arg\max_y P(y) \cdot P(x|y)$$

$$x|y=1 \sim N(\mu_1, \Sigma)$$

$$x|y=0 \sim N(\mu_2, \Sigma)$$

$$y \sim \text{Bernoulli}(\phi) \Rightarrow \begin{array}{c|cc} y & 1 & 0 \\ \hline P & \phi & 1-\phi \end{array} \quad \begin{cases} \phi^y & y=1 \\ (1-\phi)^{1-y} & y=0 \end{cases}$$

log_likelihood:

$$L(\theta) = \log \prod_{i=1}^{N} P(x_i, y_i) = \sum_{i=1}^{N} \log(P(x_i|y_i) P(y_i))$$

$$= \sum_{i=1}^{N} [\log P(x_i|y_i) + \log P(y_i)]$$

$$= \sum_{i=1}^{N} [\log N(\mu_1, \Sigma)^{y_i} + \log N(\mu_2, \Sigma)^{1-y_i} + \log \phi^{y_i} (1-\phi)^{1-y_i}]$$

Naive Bayes Classifier

最简单的有向概率图

$$\hat{y} = \arg\max_y P(y|x)$$

$$= \arg\max_{y=\{0,1\}} \frac{P(x,y)}{P(x)} = \arg\max_{y=\{0,1\}} P(x,y)$$

$$= \arg\max_y P(y) \cdot P(x|y)$$

条件独立性假设：$P(x|y) = \prod_{j=1}^{p} P(x_j|y)$

$\begin{cases} x\text{离散} \rightarrow x_j \sim \text{Categorical Dist} \\ x\text{连续} \rightarrow x_j \sim N(\mu_j, \sigma_j^2) \end{cases}$

(五) 降维. (协方差矩阵计算的协方差是feature之间的协方差)

过拟合 $\begin{cases} \text{data} \uparrow \\ \text{正则化} \\ \text{降维} \rightarrow \begin{cases} \text{直接降维(特征选择)[lasso]} \\ \text{线性降维} \\ \text{非线性降维} \end{cases} \end{cases}$

Sample Mean: $\bar{x}_{p\times 1} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}(x_1 \cdots x_N)\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}_{N\times 1} = \frac{1}{N}X^T 1_N$

Sample Covariance: $S_{p\times p} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(x_i-\bar{x})^T$

$$S = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(x_i-\bar{x})^T = \frac{1}{N}X^T H X$$

$$= \frac{1}{N}\sum_{i=1}^{N}(x_i - \frac{1}{N}X^T 1_N)(x_i - \frac{1}{N}X^T 1_N)^T$$

$$= \frac{1}{N}\underbrace{(x_1-\bar{x}, x_2-\bar{x}, \cdots, x_N-\bar{x})}_{(x_1,x_2,\cdots x_N)-(\bar{x},\bar{x},\cdots\bar{x})}\begin{pmatrix}(x_1-\bar{x})^T\\ (x_2-\bar{x})^T\\ \vdots \\ (x_N-\bar{x})^T\end{pmatrix}$$

$$= X^T - \bar{x}(1,1,\cdots 1) = X^T - \bar{x}1_N^T = X^T - \frac{1}{N}X^T 1_N 1_N^T$$

$$= \frac{1}{N}X^T\underbrace{(I_N - \frac{1}{N}1_N 1_N^T)}_{H_N \text{ centering matrix}}\cdot(I_N - \frac{1}{N}1_N 1_N^T)^T X$$

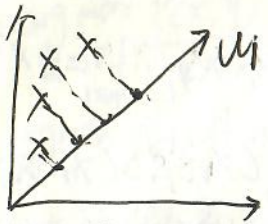
$$= \frac{1}{N}X^T H \cdot H^T X \Rightarrow \frac{1}{N}X^T H X$$

$H^2 = H\cdot H = (I_N - \frac{1}{N}1_N 1_N^T)(I_N - \frac{1}{N}1_N 1_N^T)$ $\quad H^T = H$

一个中心：原始特征空间的重构

相关 → 无关.

两个基本点：最大投影方差
最小重构距离



u_1 最大化方差的投影向量.

(中心化后均值已经为0).

$$J=\frac{1}{N}\sum_{i=1}^{N}\left(u_1^T(x_i-\bar{x})\right)^2 \quad \text{s.t. } u_1^Tu_1=1$$

(中心化指归化)

$$=\sum_{i=1}^{N}\frac{1}{N}u_1^T(x_i-\bar{x})(x_i-\bar{x})^Tu_1$$

$$=u_1^TSu_1$$

$$\begin{cases}\hat{u}_1=\arg\max u_1^TSu_1\\ \text{s.t. } u_1^Tu=1\end{cases}$$

拉格朗日项 $\lambda(1-u_1^Tu)$

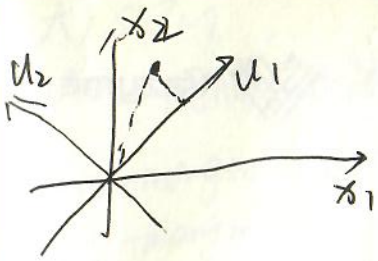
$$L(u,\lambda)=u_1^TSu_1+\lambda(1-u_1^Tu)$$

$$\frac{\partial L}{\partial u_1}=2Su_1+\lambda(-2u)=0$$

$$Su_1=\lambda u_1$$

u_1 为 S 的 eigen-vector

λ 为 S 的 eigen-value.



原数据 $x_i=\sum_{k=1}^{P}(x_i^Tu_k)u_k$

(坐标值 向量)

$$\hat{x}_i=\sum_{k=1}^{q}(x_i^Tu_k)u_k \quad \text{只用了前 } q \text{ 组 } q<P$$

$$J=\frac{1}{N}\sum_{i=1}^{N}\|x_i-\hat{x}_i\|^2 \quad \text{(前 } q \text{ 维相同)}$$

$$=\frac{1}{N}\sum_{i=1}^{N}\left\|\sum_{k=q+1}^{P}(x_i^Tu_k)u_k\right\|^2$$

$$=\frac{1}{N}\sum_{i=1}^{N}\sum_{k=q+1}^{P}\left[(x_i^Tu_k)u_k\right]^2$$

(坐标值 坐标)

$$\triangleq\frac{1}{N}\sum_{i=1}^{N}\sum_{k=q+1}^{P}\left((x_i-\bar{x})^Tu_k\right)^2$$

$$=\sum_{i=1}^{N}\frac{1}{N}\sum_{k=q+1}^{P}\left((x_i-\bar{x})^Tu_k\right)^2$$

$$=\sum_{k=q+1}^{P}\sum_{i=1}^{N}\frac{1}{N}\left((x_i-\bar{x})^Tu_k\right)^2$$

$$\left(u_k^TSu_k\right)$$

$$=\sum_{k=q+1}^{P}u_k^TSu_k \quad \text{s.t. } u_k^Tu_k=1$$

$$\begin{cases}u_k=\arg\min\sum_{k=q+1}^{P}u_k^TSu_k\\ \text{s.t. } u_k^Tu_k=1\end{cases}$$

减去的特征值最小化 = 留下的特征值最大.

SVD for PCA.

$HX = U\Sigma V^T$ （标准化）

$\begin{cases} U^TU = I \\ V^TV = VV^T = I \\ \Sigma \text{ (对角)} \end{cases}$

* $S = X^THX = X^TH^THX = V\Sigma U^TU\Sigma V^T = V\Sigma^2V^T$ (P×P)

$T = HXX^TH = U\Sigma V^TV\Sigma U^T = U\Sigma^2U^T$ (N×N)

S: 特征分解，得到方向（主成分），然后 $HX\cdot V \to$ 坐标

T: 特征分解，直接得到坐标 $HX = U\Sigma V$

$HX\cdot V = U\Sigma V^TV = U\Sigma \to$ 坐标矩阵.

$$X = (x_1, x_2, \cdots x_N)^T = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \begin{matrix} \} n \\ \downarrow \text{Sample} \end{matrix} \Bigg| \underbrace{\begin{pmatrix} x_{11} \cdots x_{1p} \\ x_{21} \cdots x_{2p} \\ \vdots \\ x_{n1} \cdots x_{np} \end{pmatrix}}_{p \to \text{feature}}$$

$HX = \underbrace{U\Sigma}_{\text{坐标矩阵}} \underbrace{V}_{\text{主成分方向}}$

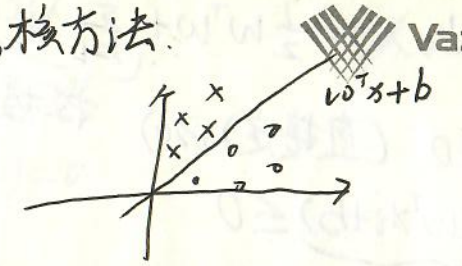
(六). SVM

SVM：间隔，对偶，核方法.

hard-margin SVM

soft-margin SVM

Kernel SVM



最大间隔分类器

max margin (w,b)

s.t. $y_i(w^Tx_i+b) > 0$

(for $\forall i = 1,2,\cdots N$)

点到直线的距离公式：

$\min_{w,b,x_i} \frac{1}{\|w\|} |w^Tx_i + b|$

$i = 1 \cdots N$

最大间隔 $\Rightarrow \max_{w,b} \min_{x_i} \frac{1}{\|w\|} y_i(w^Tx_i+b)$

$$\begin{cases} = \max_{w,b} \frac{1}{\|w\|} \left[\min_{x_i} y_i(w^Tx_i+b)\right] \\ \text{s.t. } y_i(w^Tx_i+b) > 0 \Rightarrow \exists r > 0, \min_{x_i,y_i} y_i(w^Tx_i+b) = r \end{cases}$$

由于 w^Tx_i+b 可以任意缩放 $n\cdot w^Tx_i + n\cdot b$，设 $r = 1$

$$\Rightarrow \begin{cases} \max_{w,b} \frac{1}{\|w\|} \\ \text{s.t. } \min y_i(w^Tx_i+b) = 1 \to y_i(w^Tx_i+b) \geq 1, \ i = 1,2,\cdots N. \end{cases}$$

$$\Rightarrow \begin{cases} \min_{w,b} \frac{1}{2} w^Tw \\ \text{s.t. } y_i(w^Tx_i+b) \geq 1 \quad \text{for } \forall i = 1,2,\cdots N. \end{cases}$$ (凸优化问题)

找到对偶问题 dual problem.

$L(w,b,\lambda)=\frac{1}{2}w^Tw+\sum_{i=1}^{N}\lambda_i(1-y_i(w^Tx_i+b))$

拉格朗日项.

* $\lambda_i\geq 0$ (直接定$\lambda_i\geq 0$)

$1-y_i(w^Tx_i+b)\leq 0$

$\max_{\lambda} L(w,b,\lambda)=\frac{1}{2}w^Tw+0$

$\min_{w,b}\frac{1}{2}w^Tw=\min_{w,b}\max_{\lambda}L(w,b,\lambda)$

s.t. $\lambda_i\geq 0$

对偶问题 $\begin{cases}\max_{\lambda}\min_{w,b}L(w,b,\lambda)\\ s.t.\ \lambda_i\geq 0\end{cases}$

定理: $\min\max f(x)\geq\max\min f(x)$ (弱对偶)

$\min\max f(x)=\max\min f(x)$ 强对偶.

这里的条件满足强对偶.

$\frac{\partial L}{\partial b}=0\Rightarrow\sum_{i=1}^{N}\lambda_i y_i=0$, 代入$L(w,b,\lambda)$

$\frac{\partial L}{\partial w}=0\Rightarrow w=\sum_{i=1}^{N}\lambda_i y_i x_i$. 代入

$\min_{w,b}L(w,b,\lambda)=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j x_i^Tx_j+\sum_{i=1}^{N}\lambda_i$

$\max_{\lambda}$ →

KKT条件

$$\begin{cases}\frac{\partial L}{\partial w}=0\ \ \frac{\partial L}{\partial b}=0\ \ \frac{\partial L}{\partial\lambda}=0\\ \lambda_i(1-y_i(w^Tx_i+b))=0\\ \lambda_i\geq 0\\ 1-y_i(w^Tx_i+b)\leq 0\end{cases}$$

$w^*=\sum_{i=0}^{N}\lambda_i y_i x_i$ → (只有 $w^Tx_i+b=1$或-1, $\lambda_i\neq 0$)

$\exists(x_k,y_k)$, $w^Tx+b=1\Rightarrow 1-y_k(w^Tx_k+b)=0$

$y_k(w^Tx_k+b)=1$

$y_k^2(w^Tx_k+b)=y_k$ ($y_k^2=1$)

$b^*=y_k-w^Tx_k=y_k-\sum_{i=0}^{N}\lambda_i y_i x_i^Tx_k$

$w^*=\sum_{i=0}^{N}\lambda_i y_i x_i$ (data线性组合)由支撑向量决定

$b^*=y_k-\sum_{i=0}^{N}\lambda_i y_i x_i^Tx_k$

$f(x)=sign(w^{*T}x+b^*)$

soft-margin SVM.

$\min \frac{1}{2}w^Tw + loss$

loss用距离表示

如果 $y_i(w^Tx_i+b) \geq 1$, $loss = 0$

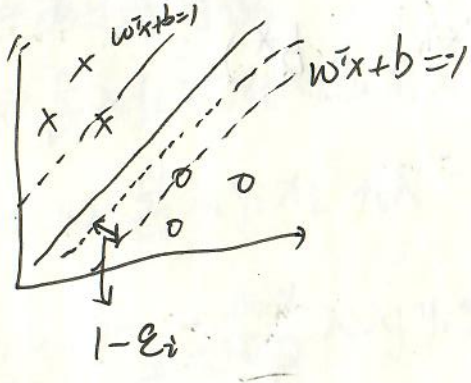
如果 $y_i(w^Tx_i+b) < 1$, $loss = 1 - y_i(w^Tx_i+b)$

$loss = \max\{0, 1-y_i(w^Tx_i+b)\}$

$$\begin{cases} \min\limits_{w,b} \frac{1}{2}w^Tw + C\sum\limits_{i=1}^{N} \max\{0, 1-y_i(w^Tx_i+b)\} \\ s.t.\ y_i(w^Tx_i+b) \geq 1-\xi_i \end{cases}$$

$$\begin{cases} \min\limits_{w,b} \frac{1}{2}w^Tw + C\sum\limits_{i=1}^{N} \xi_i \\ s.t.\ y_i(w^Tx_i+b) \geq 1-\xi_i \\ \xi_i \geq 0 \end{cases}$$

$\xi_i = \max(0, 1-y_i(w^Tx_i+b))$



约束优化问题. (原问题) I

$$\begin{cases} \min f(x) \quad x \in \mathbb{R}^p \\ s.t.\ m_i(x) \leq 0 \quad i=1,\dots,m \\ \quad\ n_j(x) = 0 \quad j=1,\dots,n \end{cases} \quad ①$$

拉格朗日函数 (原问题拉格朗日化) II

$L(x,\lambda,\eta) = f(x) + \sum\limits_{i=1}^{m}\lambda_i m_i + \sum\limits_{j=1}^{n}\eta_j n_j$

$$\begin{cases} \min\limits_{x}\max\limits_{\lambda,\eta} L(x,\lambda,\eta) \\ s.t.\ \lambda_i \geq 0 \end{cases} \quad ②$$

①与②等价

如果x违反约束 $m_i(x)$, $m_i(x)>0$, $\max\limits_{\lambda} L \to \infty$

如果x符合 $m_i(x) \leq 0$, $\max\limits_{\lambda} L \neq +\infty$

$$\min_x \max_\lambda L = \min_x \{\max_\lambda L, +\infty\} = \min_x \max_\lambda L$$

对偶问题: III

$$\begin{cases} \max\limits_{\lambda,\eta} \min\limits_{x} L(x,\lambda,\eta) \\ s.t.\ \lambda_i \geq 0 \end{cases}$$

$\max\limits_{\lambda,\eta}\min\limits_{x} L(x,\lambda,\eta) \leq \min\limits_{x}\max\limits_{\lambda,\eta} L(x,\lambda,\eta)$

弱对偶性: III ≤ II

证明: $\max\limits_{\lambda,\eta}\min\limits_{x} L(x,\lambda,\eta) \leq \min\limits_{x}\max\limits_{\lambda,\eta} L(x,\lambda,\eta)$

$\min\limits_{x} L(x,\lambda,\eta) \leq L(x,\lambda,\eta) \leq \max\limits_{\lambda,\eta} L(x,\lambda,\eta)$

对偶性的几何解释

$$\begin{cases} \min\limits_{x\in\mathbb{R}^p} f(x) \\ \text{s.t. } m_i(x)\le 0 \end{cases} \qquad D: \text{定义域} \quad D = \text{dom} f \cap \text{dom} m_i$$

$L(x,\lambda) = f(x) + \lambda m_i(x)$

$p^* = \min f(x)$ (原问题最优解)

$d^* = \max\limits_{\lambda}\min\limits_{x} L(x,\lambda)$ (对偶问题最优解)

集合 G_1: $G_1 = \{(m_i(x), f(x)) \mid x\in D\}$

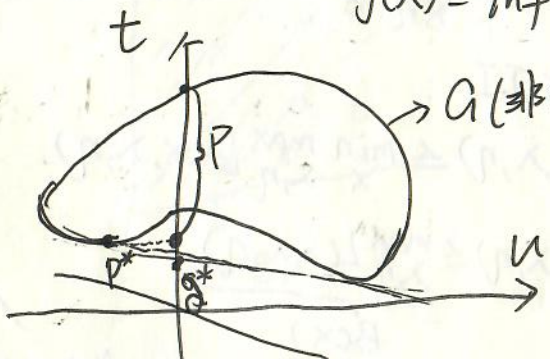
$= \{(u,t) \mid x\in D\}$

$p^* = \inf\{t \mid (u,t)\in G_1, u\le 0\}$ inf: 下确界

$d^* = \max\limits_{\lambda}\min\limits_{x} L(x,\lambda) = \max\limits_{\lambda} \underbrace{\min\limits_{x}(t+\lambda u)}_{g(\lambda)}$

$= \max\limits_{\lambda} g(\lambda)$

$g(\lambda) = \inf\{t+\lambda u \mid (u,t)\in G_1\}$



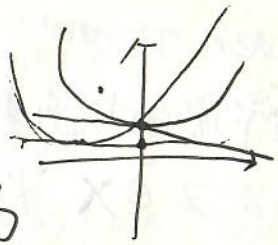
$d^* < p^*$

Slater Condition

$\exists \hat{x} \in \text{relint } D$ (relative interior)

s.t. $\forall i = 1,\cdots,m \quad m_i(\hat{x}) < 0$

① 对于大多数凸优化，slater成立



KKT condition

凸优化 + slater $\Rightarrow$ 强对偶

$\Updownarrow$

KKT.

KKT:

$$\begin{cases} \text{可行条件}: \begin{cases} m_i(x^*) \le 0 \\ n_j(x^*) = 0 \\ \lambda^* \ge 0 \end{cases} \\ \text{互补松弛}: \lambda_i m_i = 0 \\ \text{梯度为0}: \left.\dfrac{\partial f(x,\lambda^*,\eta^*)}{\partial x}\right|_{x=x^*} \end{cases}$$

$d^* = \max\limits_{\lambda,\eta} g(\lambda,\eta) = g(\lambda^*,\eta^*)$

$= \min\limits_{x} L(x,\lambda^*,\eta^*)$

$\le_{\Delta} L(x^*,\lambda^*,\eta^*)$

$= f(x^*) + \sum \lambda_i^* m_i + \sum \eta_j^* n_j$

$\le_{\Delta} f(x^*)$

$= p^*$

Kernel Method. 核方法 (七)

Vazyme

非线性带来高维转换 (从模型角度)
对偶表示带来内积 (从优化角度).

PLA(Perceptron Learning Algorithm) → 多层感知机 (神经网络)

→ 深度学习

cover Theorem: 高维比低维更易线性可分.

通常说的核函数是正定核函数.

$\forall x, z \in \mathcal{X}$, $K(x,z) = \langle \phi(x) \cdot \phi(z) \rangle$ 内积

$K: \mathcal{X} \cdot \mathcal{X} \mapsto \mathbb{R}$ 核函数

正定核: $K: \mathcal{X} \cdot \mathcal{X} \mapsto \mathbb{R}$

$\forall x, z \in \mathcal{X}$, 有 $K(x,z)$

如果 $\exists: \phi: \mathcal{X} \mapsto \mathbb{R}$, $\phi \in \mathcal{H}$

s.t. $K(x,z) = \langle \phi(x), \phi(z) \rangle$, 则 $K(x,z)$ 为正定核函数

正定核: $K: \mathcal{X} \cdot \mathcal{X} \mapsto \mathbb{R}$, $\forall x, z \in \mathcal{X}$, 有 $K(x,z)$

如果 $K(x,z)$ 满足如下性质:

① 对称性 ② 正定性

那么称 $K(x,z)$ 为正定核函数.

① 对称性 $\Leftrightarrow K(x,z) = K(z,x)$

要证: $K(x,z) = \langle \phi(x), \phi(z) \rangle \Leftrightarrow$ Gram matrix 半正定.

Vazyme

半正定.

Hilbert Space:

完备的, 可能是无限维的, 被赋予内积的线性空间.

向量空间

→ 对极限是封闭的

$\{K_n\}$ $\lim_{n\to\infty} K_n = K \in \mathcal{H}$

→ 对称性
正定性
线性
- $\langle f, g \rangle = \langle g, f \rangle$
- $\langle f, f \rangle \geq 0$
- $\langle r_1 f_1 + r_2 f_2, g \rangle = r_1 \langle f_1, g \rangle + r_2 \langle f_2, g \rangle$

必要性:

$K(x,z) \Rightarrow \langle \phi(x), \phi(z) \rangle \Rightarrow$ Gram matrix 半正定.

对称: $K(x,z) = \langle \phi(x), \phi(z) \rangle = \langle \phi(z), \phi(x) \rangle = K(z,x)$

正定: $\begin{pmatrix} K(x_1,x_1) & \cdots & K(x_1,x_N) \\ \vdots & \ddots & \vdots \\ K(x_N,x_1) & \cdots & K(x_N,x_N) \end{pmatrix}$

↓

证 $\forall \lambda \in \mathbb{R}^N$, $\lambda^T K \lambda \geq 0$

$\lambda^T K \lambda = (\lambda_1, \lambda_2, \cdots \lambda_N) \begin{pmatrix} K_{11} & K_{12} & \cdots & K_{1N} \\ K_{21} & & & \vdots \\ \vdots & & & \vdots \\ K_{N1} & \cdots & \cdots & K_{NN} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \end{pmatrix}$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j K_{ij}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j \langle \phi(x_i), \phi(x_j) \rangle$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j \phi(x_i)^T \phi(x_j)$$

$$= \sum_{i=1}^{N} a_i \phi(x_i)^T \cdot \sum_{j=1}^{N} a_j \phi(x_j)$$

$$= \left[\sum_{i=1}^{N} a_i \phi(x_i)\right]^T \cdot \sum_{j=1}^{N} a_j \phi(x_j)$$

$$= \left\langle \sum_{i=1}^{N} a_i \phi(x_i), \sum_{j=1}^{N} a_j \phi(x_j) \right\rangle$$

$$= \left\| \sum_{i=1}^{N} a_i \phi(x_i) \right\|^2 \geq 0.$$

(八) 指数族分布.

二项分布
负二项分布
泊松分布
几何分布
指数分布

Beta
Dirichlet
Gamma.
Gaussian (高斯分布/正态分布)

指数族分布标准型式: $P(x|\eta) = h(x)\exp(\eta^T \phi(x) - A(\eta))$

η: 参数向量, $x \in \mathbb{R}^p$

$A(\eta)$ log partition function.
对数. 配分函数.

$A'(\eta) = E[\phi(x)]$

$A''(\eta) = \mathrm{Var}[\phi(x)]$

充分统计量: 能够表示分布的统计量(均值, 方差...)
$\phi(x)$

共轭: 后验 $\propto$ 似然 $\times$ 先验

$P(z|x) \propto P(x|z)P(z)$

当似然为指数族分布, 先验的分布型式与后验相同
计算上的方便.

广义线性模型: 线性组合 $\eta^T x$
link function → (激活函数)$^{-1}$

极大似然估计与充分统计量.

$D=\{x_1, x_2, \cdots, x_N\}$

$\eta_{MLE} = \arg\max \log P(D|\eta)$

$= \arg\max \log \prod_{i=1}^{N} P(x_i|\eta)$

$= \arg\max \sum_{i=1}^{N} (\eta^T \phi(x_i) - A(\eta))$

$\frac{\partial}{\partial \eta}\left(\sum_{i=1}^{N}(\eta^T\phi(x_i) - A(\eta)\right)$

$= \sum_{i=1}^{N} \phi(x_i) - A'(\eta) = 0 \quad A'(\eta_{MLE}) = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)$

最大熵:

信息量: $-\log P$

熵: $E_{P(x)}[-\log P] = \int -p(x)\cdot \log p(x)\, dx$

$= -\sum_x p(x)\cdot \log P(x)$

$$\begin{cases} \max H[P] = \max\left(-\sum_{i=1}^{K} P_i \log P_i\right) \\ s.t. \sum_{i=1}^{K} P_i = 1 \end{cases} \Rightarrow \begin{cases} \min \sum_{i=1}^{K} P_i \log P_i \\ s.t. \sum_{i=1}^{K} P_i = 1 \end{cases}$$

$L(P,\lambda) = \sum_{i=1}^{K}(P_i \log P_i) + \lambda\left(1 - \sum_{i=1}^{K} P_i\right)$

$\hat{P}_i = \exp(\lambda - 1)$

$\because \hat{P}_1 = \hat{P}_2 = \cdots = \hat{P}_K$.

$\therefore P(x)$ 均匀分布

当经验分布已知:

$f(x)$ 是关于 x 的函数

$E_P[f(x)] = \Delta$ (已知)

$$\begin{cases} \min \sum_x P(x)\log P(x) \\ s.t. \sum_x P(x) = 1 \\ \cdots E_P[f(x)] = \Delta \end{cases}$$

(九) 概率图.

高维随机变量 $P(x_1, x_2, \cdots x_p)$ {边缘概率: $P(x_i)$; 条件概率: $P(x_j|x_i)$}

Sum Rule: $P(x_1) = \int P(x_1, x_2)\, dx_2$

Product Rule: $P(x_1, x_2) = P(x_1)P(x_2|x_1) = P(x_2)\cdot P(x_1|x_2)$

Chain Rule: $P(x_1, x_2, \cdots x_p) = \prod_{i=1}^{p} P(x_i|x_1, x_2, \cdots x_{i-1})$

Bayesian Rule: $P(x_2|x_1) = \frac{P(x_1,x_2)}{P(x_1)} = \frac{P(x_1,x_2)}{\int P(x_1,x_2)dx_2} = \frac{P(x_2)P(x_1|x_2)}{\int P(x_2)P(x_1|x_2)dx_2}$

维度高, 计算复杂

↓

简化 → 相互独立 → $P(x_1, x_2, \cdots x_p) = \prod_{i=1}^{p} P(x_i)$

↳ Naive Bayes: $P(x|y) = \prod_{i=1}^{p} P(x_i|y)$

Markov Property → $x_j \perp x_{j+1} | x_i, j<i$ (x_i 与 x_{j+1} 未次, 同构)

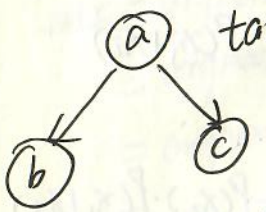
条件独立性 → $x_A \perp x_B | x_C$

概率图 {
- 表示 — 有向图 Bayesian Network; 无向图 Markov Network
- 推断 — 精确推断; 近似推断 — 确定性近似(变为推断); 随机近似(MCMC)
- 学习 — (参数学习 — 完备数据)

Bayesian Network

因子分解 $P(x_1, x_2, \cdots, x_p) = \prod_{i=1}^{p} P(x_i \mid x_{pa(i)})$

x_i的父亲集合.



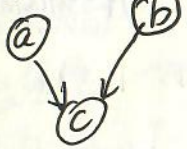
tail to tail

$c \perp b \mid a$

head to tail

$a \to b \to c$ $\quad a \perp c \mid b$

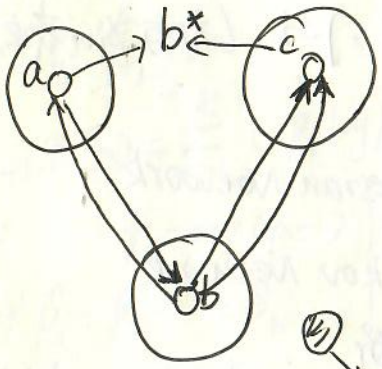
head to head



若C未被观测 $a \perp b$

若C被观测,则a与b不独立.

D-seperation.



$$P(x_i \mid x_{-i}) = \frac{P(x_i, x_{-i})}{P(x_{-i})}$$

挑选与x_i相关的

$$= \frac{P(x)}{\int_{-i} P(x)\, dx_i}$$

EM (Expectation Maximum) 期望最大.

MLE: $P(x \mid \theta)$ $\quad \theta_{MLE} = \arg\max_{\theta} \log P(x \mid \theta)$

$$\theta^{t+1} = \arg\max_{\theta} \int_z \underbrace{\log P(x, z \mid \theta) \cdot P(z \mid x, \theta^{(t)})}_{E_{z \mid x, \theta^{(t)}}[\log P(x, z \mid \theta)]} dz$$

$\theta^t \to \theta^{t+1}$

$\log P(x \mid \theta^t) \le \log P(x \mid \theta^{t+1})$

$$\log P(x \mid \theta) = \log P(x, z \mid \theta) - \log P(z \mid x, \theta)$$

① ②

$$① = \log P(x \mid \theta) \underbrace{\int_z P(z \mid x, \theta^{(t)})\, dz}_{1} = \log P(x \mid \theta)$$

$$② = \underbrace{\int_z P(z \mid x, \theta^t) \cdot \log P(x, z \mid \theta)\, dz}_{Q(\theta, \theta^t)} - \underbrace{\int_z P(z \mid x, \theta^t) \cdot \log P(z \mid x, \theta)\, dz}_{H(\theta, \theta^t)}$$

$Q(\theta^{t+1}, \theta^t) \ge Q(\theta^t, \theta^t)$

$H(\theta^{t+1}, \theta^t) \le H(\theta^t, \theta^t)$

$H(\theta^{t+1}, \theta^t) - H(\theta^t, \theta^t) \le 0$

收敛性: $\log P(x \mid \theta^{t+1}) \ge \log P(x \mid \theta^t)$

$\log P(x|\theta) = \log P(x,z|\theta) - \log P(z|x,\theta)$

$P(x,z) = P(z|x)P(x)$

$\log P(x) = \log \frac{P(x,z)}{P(z|x)} = \log P(x,z) - \log P(z|x)$

$\log P(x|\theta) = \log P(x,z|\theta) - \log P(z|x,\theta)$

交迭两

$= \log \frac{P(x,z|\theta)}{q(z)} - \log \frac{P(z|x,\theta)}{q(z)}$ $q(z) \neq 0$ ($q(z)$关于z的分布函数).

左边 $= \int_z q(z) \cdot \log P(x|\theta)\, dz = \log P(x|\theta) \cdot \int_z q(z)\, dz = \log P(x|\theta)$

右边 $= \int_z q(z) \log \frac{P(x,z|\theta)}{q(z)} dz - \int_z q(z) \log \frac{P(z|x,\theta)}{q(z)} dz$

ELBO (evidence lower Bound) $KL(q(z) \| P(z|x,\theta))$

$\log P(x|\theta) = ELBO + KL(q\|P)$

($KL(q\|P) \geqslant 0$)

$\log P(x|\theta) \uparrow \geqslant ELBO \uparrow$

$\theta = \arg\max_\theta ELBO = \arg\max_\theta \int q(z) \log \frac{P(x,z|\theta)}{q(z)} dz$

$= \arg\max_\theta \int P(z|x,\theta^t) \log \frac{P(x,z|\theta)}{P(z|x,\theta^t)} dz$

$= \arg\max_\theta \int P(z|x,\theta^t) \cdot \log P(x,z|\theta) dz$

$\hat{q} = P(z|x,\theta^t)$

E-step: $P(z|x,\theta^t) \to E_{z|x,\theta^t}[\log P(x,z|\theta)]$

M-step $\theta^{t+1} = \arg\max_\theta E_{z|x,\theta^t}[\log P(x,z|\theta)]$

$\log P(x|\theta) = \log \int_z P(x,z|\theta)\, dz = \log \int_z \frac{P(x,z|\theta)}{q(z)} \cdot q(z)\, dz$

$= \log E_{q(z)}\left[\frac{P(x,z|\theta)}{q(z)}\right]$

$\geqslant E_{q(z)}\left[\log \frac{P(x,z|\theta)}{q(z)}\right]$

$=$ ELBO $\cdot$ $P(z|x,\theta^t)$

z为隐变量.

EM Algorithm $\hat{\theta} = \arg\max_\theta P(x|\theta) = \arg\max_\theta \log P(x|\theta)$

$\log P(x|\theta) = ELBO + KL(q\|p)$

$\geqslant ELBO$

$L(q,\theta)$

$q(z)$可能取不到$P(z|x,\theta)$

θ固定时, $\log P(x|\theta)$固定, $ELBO \uparrow$, $KL(q\|p) \downarrow$

相当于求 $\arg\min_q KL(q\|p) = \arg\max L(q,\theta)$

固定 $\hat{q}$, $\theta = \arg\max_\theta L(\hat{q},\theta)$

$$L(q,\theta) = E_q[\log P(x,z) - \log q]$$
$$= E_q[\log P(x,z)] - \underbrace{E_q[\log q]}_{H[q]}$$

~~MCMC: Markov Chain and Monte Carlo~~

(十二) 变分推断

Variational Inference (变分推断)

矩阵角度优化问题
1. 回归: $f(w) = w^T x$ (模型)
   loss function: $L(w) = \sum_{i=1}^{N} \| w^T x_i - y_i \|$
   $\hat{w} = \arg\min_w L(w)$ (策略)
   解法: ① 解析向下, $\frac{\partial L(w)}{\partial w} = 0 \Rightarrow w^* = (X^T X)^{-1} X^T Y$
   ② 梯度下降: GD
2. 分类
   SVM
3. EM

Inference
- 精确推断
- 近似推断
  - 确定性近似 → VI
  - 随机近似 → MCMC, MH, Gibbs.

贝叶斯角度 → 积分问题

$$P(\theta|x) = \frac{P(\theta)P(x|\theta)}{P(x)} \rightarrow \int_\theta P(x|\theta)P(\theta)\,d\theta$$

贝叶斯 inference (推断)

贝叶斯决策: 对新样本预测 $P(\tilde{x}|X)$.

$$P(\tilde{x}|X) = \int P(\tilde{x}|\theta)P(\theta|X)\,d\theta = \int P(\tilde{x}|\theta)\frac{P(\theta)P(X|\theta)}{P(X)}\,d\theta$$

X: observed data

Z: latent variable + parameter.

变为推断为了逼近后验分布 $P(\Theta|X)$.

(X, Z): complete data.

$$\log P(X) = \log P(X,Z) - \log P(Z|X)$$

$$= \log \frac{P(X,Z)}{q(Z)} - \log \frac{P(Z|X)}{q(Z)}$$

$$\int_Z q(Z) \log P(X)\, dZ = \log P(X) = \int_Z q(Z) \log \frac{P(X,Z)}{q(Z)} - \int_Z q(Z) \log \frac{P(Z|X)}{q(Z)}$$

$$= \underbrace{\int_Z q(Z) \log \frac{P(X,Z)}{q(Z)}}_{ELBO} - \underbrace{\left[\int_Z q(Z) \log \frac{P(Z|X)}{q(Z)}\right]}_{KL(q\|P)}$$

$$= L(q) + KL(q\|P)$$

$L(q)$ ↓ 变为

$KL(q\|P) \neq 0$

要使 $q(Z) \simeq P(Z|X)$

$$\Downarrow$$

$$\arg\max_{q(Z)} L(q)$$

$$q(Z) = \arg\max_{q(Z)} L(q) \Rightarrow q(Z) \simeq P(Z|X)$$

为成 M 个相互独立的组.

$$q(Z) = \prod_{i=1}^{M} q_i(Z_i)$$

$$L(q) = \int_Z q(Z) \log \frac{P(X,Z)}{q(Z)}$$

$$= \underbrace{\int_Z q(Z) \log P(X,Z)}_{①} - \underbrace{\int_Z q(Z) \log q(Z)}_{②}$$

$$① = \int_Z \prod_{i=1}^{M} q_i(Z_i) \log P(X,Z)\, dZ_1 dZ_2 \cdots dZ_M$$

$$= \int_{Z_j} q_j(Z_j) \underbrace{\left[\int_{(i\neq j)} \prod_{i\neq j}^{M} q_i(Z_i) \log P(X,Z)\, dZ_1 \cdots dZ_M\right]}_{\int_{Z_1 \cdots Z_M (i\neq j)} \log P(X,Z) \cdot \prod_{i\neq j}^{M} q_i(Z_i)\, dZ_i} dZ_j$$

$$= \int_{Z_j} q_j(Z_j) \cdot \underbrace{E_{\prod_{i\neq j} q_i(Z_i)}\left[\log P(X,Z)\right]}_{\log \tilde{P}(X,Z_j)} dZ_j$$

$$② \int_Z q(Z) \log q(Z)\, dZ = \sum_{i=1}^{M} \int_{Z_i} q_i(Z_i) \log q_i(Z_i)\, dZ_i$$

$$= \int_{Z_j} q_j(Z_j) \log q_j(Z_j)\, dZ_j + C$$

$$① - ② = \int_{Z_j} q_j(Z_j) \cdot \log \frac{\tilde{P}(X,Z_j)}{q_j(Z_j)}\, dZ_j = -KL(q_j\|\tilde{P}(X,Z_j)) + C$$

(十三) MCMC (采样方法,为了近似后验).

推断后验 $P(\theta|X)$

近似推断 {确定性→VI
随机性→MCMC.

Monte Carlo method: 基于采样的随机近似法.

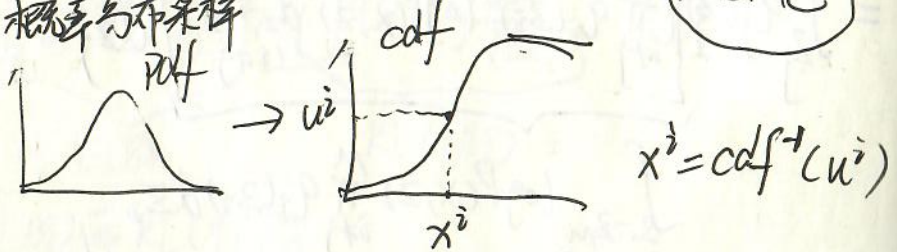
latent → observed data

$P(Z|X) \longrightarrow E_{Z|X}[f(Z)] = \int p(Z|X)\cdot f(Z)\,dZ$

$\simeq \frac{1}{N}\sum_{i=1}^{N} f(Z_i)$ ← MCMC

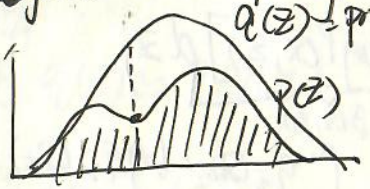
$Z^1, Z^2, \cdots Z^N \sim P(Z|X)$

① 概率分布采样



$x^i = cdf^{-1}(u^i)$

② Rejection Sampling



$\forall Z_i,\ Mq(Z_i) \geq P(Z_i)$

取样,阴影部分接收, α: 接收率 $\alpha = \frac{P(Z^i)}{Mq(Z^i)}$, $0 \leq \alpha \leq 1$

1. $Z^i \sim q(Z)$

③ Importance Sampling

$E_{P(Z)}[f(Z)] = \int_Z P(Z) f(Z)\,dZ = \int \frac{P(Z)}{q(Z)} \cdot q(Z) \cdot f(Z)\,dZ$

$= \int f(Z)\cdot \frac{P(Z)}{q(Z)} \cdot q(Z)\,dZ$

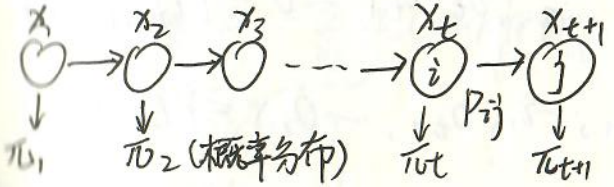
$\simeq \frac{1}{N}\sum_{i=1}^{N} f(Z^i)\frac{P(Z^i)}{q(Z^i)}$, $Z_i \sim q(Z)$ $i=1,2,\cdots N$.

$\frac{P(Z^i)}{q(Z^i)}$ → weight.

一阶(齐次) Markov chain. $P(X_{t+1}=x \mid x_1, x_2, \cdots x_t) = P(X_{t+1}=x \mid X_t)$

P → 转移矩阵 $[P_{ij}]$

$P_{ij} = P(X_{t+1}=j \mid X_t = i)$



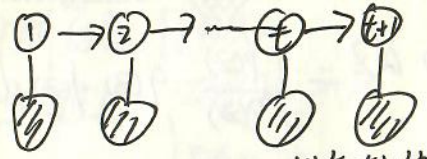
$\pi_{t+1}(x^*) = \int \pi_t(x)\cdot P(x \to x^*)\,dx$

$\pi = [\pi^{(1)}, \pi^2, \cdots \pi^N \cdots]$

$\sum_{i=1}^{\infty} \pi(i) = 1$

平稳分布: π^k 是 x_k 的平稳分布

(十四) HMM 隐马尔可夫模型.



$\lambda = (\pi, A, B)$ — A: 状态转移; B: 发散矩阵; π: 初始 prob dist

观测变量: $O_1, O_2 \cdots O_t \to V = \{v_1, v_2, \cdots, v_m\}$ (观测的值域)

隐状态变量: $i_1, i_2 \cdots i_t \to Q = \{q_1, q_2, \cdots, q_N\}$ (隐状态的值域)

$A = [a_{ij}]$, $a_{ij} = P(i_{t+1} = q_j \mid i_t = q_i)$ 状态转移矩阵.

$B = [b_j(k)]$ $b_j(k) = P(O_t = v_k \mid i_t = q_j)$ 发射矩阵, 状态 j 到观测 k

两个假设: ① 齐次 Markov 假设. ② 观察独立.

$$P(i_{t+1} \mid i_t, i_{t-1}, \cdots i_1, O_t, O_{t-1}, \cdots O_1) = P(i_{t+1} \mid i_t)$$

$$P(O_t \mid i_t, i_{t-1}, \cdots i_1, O_{t-1}, \cdots O_1) = P(O_t \mid i_t)$$

三大问题:

① Evaluation → $P(O \mid \lambda)$ 前向后向

② Learning → EM

③ Decoding $I = \arg\max P(I \mid O)$

预测 → $P(i_{t+1} \mid O_1, O_2, \cdots O_t)$

Evaluation 问题: Given λ, 求 $P(O \mid \lambda)$

$$P(O \mid \lambda) = \sum_I P(I, O \mid \lambda) = \sum_I P(O \mid I, \lambda) \cdot P(I \mid \lambda)$$

$$P(I \mid \lambda) = P(i_1, i_2, \cdots i_T \mid \lambda) = P(i_T \mid i_1, i_2, \cdots i_{T-1}, \lambda) \cdot P(i_1, i_2, \cdots i_{T-1}, \lambda)$$

$P(i_T \mid i_{T-1})$ 即 $a_{i_{T-1} i_T}$

$$= \pi(a_{i_1}) \prod_{t=2}^{T} a_{i_{t-1}, i_t}$$

$$P(O \mid I, \lambda) = \prod_{t=1}^{T} b_{i_t}(O_t)$$

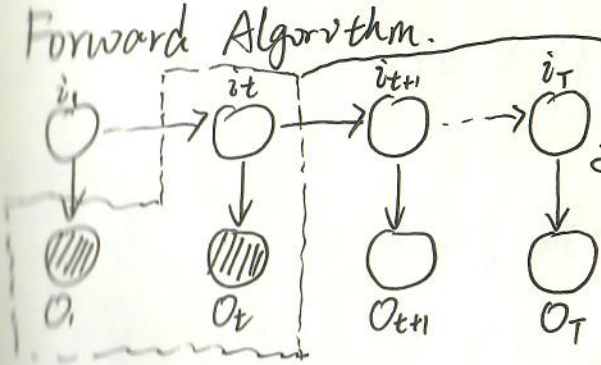
$$P(O \mid \lambda) = \sum_I \pi(a_{i_1}) \prod_{t=2}^{T} a_{i_{t-1}, i_t} \prod_{t=1}^{T} b_{i_t}(O_t)$$

$\sum_I$ I 的可能的种类为 $\sum_{i_1} \cdot \sum_{i_2} \cdots \sum_{i_t}$.

$$= \underbrace{\sum_{i_1} \sum_{i_2} \cdots \sum_{i_t}} \pi(a_{i_1}) \prod_{t=2}^{T} a_{i_{t-1}, i_t} \prod_{t=1}^{T} b_{i_t}(O_t)$$

$O(N^t)$

Forward Algorithm.



联合概率

$\alpha_t(i) = P(O_1, \cdots O_t, i_t = q_i \mid \lambda)$

$\alpha_T(i) = P(O, i_T = q_i \mid \lambda)$

$$P(O \mid \lambda) = \sum_{i=1}^{N} P(O, i_t = q_i \mid \lambda)$$

$$= \sum_{i=1}^{N} \alpha_T(i)$$

$$\alpha_{t+1}(j) = P(o_1, \cdots, o_t, o_{t+1}, i_{t+1}=q_j \mid \lambda)$$

$$= \sum_{i=1}^{N} P(o_1, \cdots, o_t, o_{t+1}, i_t=q_i, i_{t+1}=q_j \mid \lambda)$$

$$= \sum_{i=1}^{N} P(o_{t+1} \mid o_1, \sim o_t, i_t=q_i, i_{t+1}=q_j, \lambda) \cdot P(o_1, \cdots, o_t, i_t=q_i, i_{t+1}=q_j \mid \lambda)$$

$$= \sum_{i=1}^{N} P(o_{t+1} \mid i_{t+1}=q_j) \cdot P(o_1, \cdots o_t, i_t=q_i, i_{t+1}=q_j \mid \lambda)$$

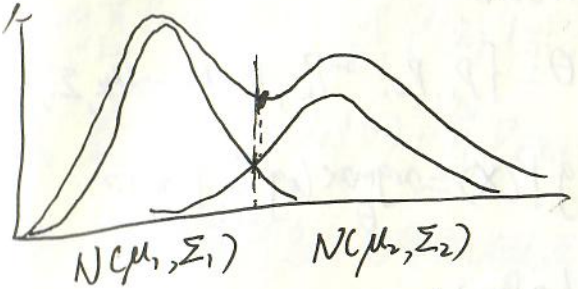
$$= \sum_{i=1}^{N} P(o_{t+1} \mid i_{t+1}=q_j) \cdot P(i_{t+1}=q_j \mid o_1, \cdots, o_t, i_t=q_i, \lambda) \cdot P(o_1, \cdots o_t, i_t=q_i \mid \lambda)$$

$$= \sum_{i=1}^{N} P(o_{t+1} \mid i_{t+1}=q_j) \cdot P(i_{t+1}=q_j \mid i_t=q_i, \lambda) \cdot \alpha_t(i)$$

$$= \sum_{i=1}^{N} b_j(o_{t+1}) \cdot a_{ij} \cdot \alpha_t(i)$$

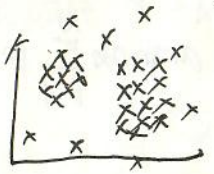
GMM: Gaussian Mixture Model (高斯混合模型)

从几何角度来看：加权平均 → (多个高斯分布叠加而成)



$$P(x) = \sum_{k=1}^{K} \alpha_k N(\mu_k, \Sigma_k), \quad \sum_{k=1}^{K} \alpha_k = 1$$

从混合模型角度来看：(生成模型)



生成N个样本.

x: observed variable

z: latent variable. (对应的样本属于哪一个高斯分布)

z为离散随机变量

z	C_1	C_2	⋯	C_K
	1	2		K
$P(z)$	P_1	P_2	⋯	P_K

$$P(x) = \sum_{z} P(x, z) = \sum_{k=1}^{K} P(x, z=C_k) = \sum_{k=1}^{K} P(z=C_k) \cdot P(x \mid z=C_k)$$

$$= \sum_{k=1}^{K} P_k \cdot N(x \mid \mu_k, \Sigma_k)$$

$$P(x, z) = P(z) \cdot P(x \mid z) = P_z \cdot N(x \mid \mu_z, \Sigma_z)$$

$$P(z \mid x) = \frac{P(x, z)}{P(x)} = \frac{P_z \cdot N(x \mid \mu_z, \Sigma_z)}{\sum_{k=1}^{K} P_k \cdot N(x \mid \mu_k, \Sigma_k)}$$

X: observed data → $X=(x_1, x_2, \cdots x_N)$

(X, Z): complete data

θ: parameter → $\theta=\{p_1, p_2, \cdots p_k, \mu_1, \mu_2 \cdots \mu_k, \Sigma_1, \Sigma_2 \cdots \Sigma_k\}$

$$\hat{\theta}_{MLE}=\arg\max_\theta \log P(X)=\arg\max_\theta \log \prod_{i=1}^N P(x_i)=$$

$$=\arg\max_\theta \sum_{i=1}^N \log P(x_i)$$

$$=\arg\max_\theta \sum_{i=1}^N \log \underbrace{\sum_{k=1}^K p_k N(x_i|\mu_k, \Sigma_k)}_{\text{混合高斯分布}}$$

无法直接得到解析解，所以用EM算法求近似解.

使用EM算法.

$$\theta^{t+1}=\arg\max_\theta \underbrace{E_{Z|X,\theta^t}[\log P(X, Z|\theta)]}_{Q(\theta, \theta^t)}$$

$$Q(\theta, \theta^t)=\int_Z \log P(X, Z|\theta) P(Z|X, \theta^t) dZ$$

$$=\sum_Z \log \prod_{i=1}^N P(x_i, z_i|\theta) \prod_{i=1}^N P(z_i|x_i, \theta^t)$$

$$=\sum_{z_1, z_2 \cdots z_N} \sum_{i=1}^N \log P(x_i, z_i|\theta) \cdot \prod_{i=1}^N P(z_i|x_i, \theta^t)$$

z_1：第一个样本属于各个分布的概率

$$=\sum_{z_1, z_2 \cdots z_N} [\log P(x_1, z_1|\theta)+\log P(x_2, z_2|\theta)+\cdots+\log P(x_N, z_N|\theta)] \prod_{i=1}^N P(z_i|x_i, \theta^t)$$

$$\sum_{z_1 \cdots z_N} \log P(x_1, z_1|\theta) \prod_{i=1}^N P(z_i|x_i, \theta^t)$$

$$=\sum_{z_1 \cdots z_N} \log P(x_1, z_1|\theta) P(z_1|x_1, \theta^t) \prod_{i=2}^N P(z_i|x_i, \theta^t)$$

$$=\sum_{z_1} \log P(x_1, z_1|\theta) P(z_1|x_1, \theta^t) \underbrace{\sum_{z_2 \cdots z_N} \prod_{i=2}^N P(z_i|x_i, \theta^t)}_{=1}$$

$$=\sum_{z_1} \log P(x_1, z_1|\theta) P(z_1|x_1, \theta^t)$$

$$=\sum_{z_1} \log P(x_1, z_1|\theta) \cdot P(z_1|x_1, \theta^t)+\cdots+$$

$$\sum_{z_N} \log P(x_N, z_N|\theta) \cdot P(z_N|x_N, \theta^t)$$

$$=\sum_{i=1}^N \sum_{z_i} \log P(x_i, z_i|\theta) \cdot P(z_i|x_i, \theta^t)$$

E-step:

$$Q(\theta,\theta^t)=\int_Z \log P(X,Z|\theta)\,P(Z|X,\theta^t)\,dZ$$

$$=\sum_{i=1}^{N}\sum_{z_i}\log P_{z_i}\cdot N(x_i|\mu_{z_i},\Sigma_{z_i})\,\frac{P_{z_i}\cdot N(x_i|\mu_{z_i}^t,\Sigma_{z_i}^t)}{\sum_{k=1}^{K}P_k^t\cdot N(x_i|\mu_k^t,\Sigma_k^t)}$$

$$=\sum_{i=1}^{N}\sum_{z_i}\log\left[P_{z_i}\cdot N(x_i|\mu_{z_i},\Sigma_{z_i})\right]P(z_i|x_i,\theta^t)$$

$$=\sum_{z_i}\sum_{i=1}^{N}\log\left[P_{z_i}N(x_i|\mu_{z_i},\Sigma_{z_i})\right]P(z_i|x_i,\theta^t)$$

$$=\sum_{k=1}^{K}\sum_{i=1}^{N}\log\left[P_k\cdot N(x_i|\mu_k,\Sigma_k)\right]\cdot P(z_i=c_k|x_i,\theta^t)$$

$$=\sum_{k=1}^{K}\sum_{i=1}^{N}\left[\log P_k+\log N(x_i|\mu_k,\Sigma_k)\right]\cdot P(z_i=c_k|x_i,\theta^t)$$

M-step:

$$P^{t+1}=(P_1^{t+1},P_2^{t+1},\cdots,P_K^{t+1})$$

$$\begin{cases}\max\limits_{P}\sum\limits_{k=1}^{K}\sum\limits_{i=1}^{N}\log P_k\cdot P(z_i=c_k|x_i,\theta^{t+1})\\ s.t.\ \sum\limits_{k=1}^{K}P_k=1\end{cases}$$

$$L(P,\lambda)=\sum_{k=1}^{K}\sum_{i=1}^{N}\log P_k\cdot P(z_i=c_k|x_i,\theta^t)+\lambda\left(\sum_{k=1}^{K}P_k-1\right)$$

$$\frac{\partial L(P,\lambda)}{\partial P_k}=\sum_{i=1}^{N}\frac{1}{P_k}P(z_i=c_k|x_i,\theta^t)+\lambda=0$$

$$\sum_{i=1}^{N}P(z_i=c_k|x_i,\theta^t)+P_k\cdot\lambda=0$$

$$\xrightarrow{k=1,\cdots,K}\sum_{i=1}^{N}\sum_{k=1}^{K}P(z_i=c_k|x_i,\theta^t)+\sum_{k=1}^{K}P_k\cdot\lambda=0$$

$$N+\lambda=0$$

$$\lambda=-N$$

$$\sum_{i=1}^{N}P(z_i=c_k|x_i,\theta^t)+P_k(-N)=0$$

$$P_k=\frac{1}{N}\sum_{i=1}^{N}P(z_i=c_k|x_i,\theta^t)$$

$$\theta=\{P_1,P_2,\cdots P_K,\mu_1,\mu_2,\cdots\mu_K,\Sigma_1,\Sigma_2,\cdots\Sigma_K\}$$

(三十) 生成模型 (关注样本分布本身) P(X)

监督
- 概率模型
  - 判别 (P(Y|X)): 逻辑回归, 最大熵
  - 生成模型
- 非概率模型: PLA, SVM, LDA, KNN, NN, Tree Model

非监督
- 概率模型: 生成模型
- 非概率模型: PCA, K-means, Auto-encoder

生成模型:

Naive Bayes 假设: $P(x|y) = \prod_{i=1}^{n} P(x_i|y)$

Mixture model: GMM

Time-series Model: HMM | KF | PF (Kalman, Particle)

非参贝叶斯模型: GP | DP

标示: 形神兼备

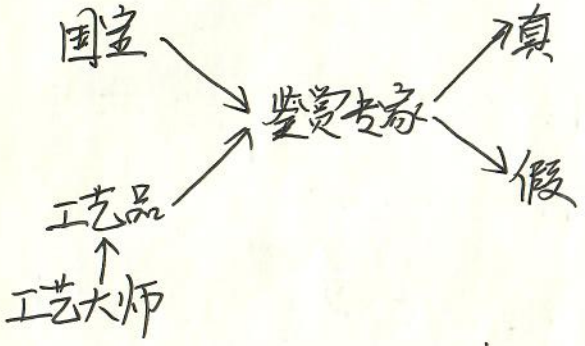
形
- Discrete vs. Continuous. (节点可以离散或连续)
- Directed Model vs. Undirected
- Latent Variable vs. Fully-observed
- shallow vs. Deep (多层次)
- Sparse vs. Dense (多连接)

神:
- parameteric vs. non-parameteric
- Implicit Density vs. Explicit Density
  - Implicit Density ⇓ GAN
  - Explicit Density ⇓ P(x)

推断: tractable vs. intractable.

学习: MLE: maximum likelihood-based vs. likelihood-free Model.

(三十一) Generative Adversarial Network. 生成对抗网络.



国宝: $P_{data}(x) : \{x_i\}_{i=1}^{N}$

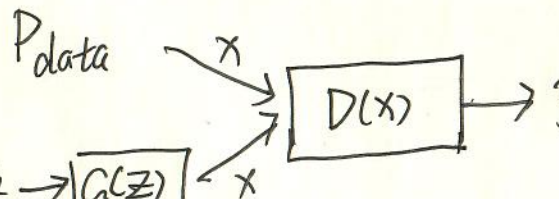
工艺品: $P_g(x;\theta_g) : generator(P_z(z) + G(z,\theta_g))$

$P_z(z)$ → 分布　　$G(z,\theta_g)$ → NN, 产生 x

鉴赏专家: $y|x$ discriminator

$y\|x$	1	0
P	$D(x)$	$1-D(x)$

目标: ①鉴赏专家鉴别能力强. (手段)
②工艺大师制造能力强 (目标)



$$\min_G \max_D E_{x\sim P_{data}}[\log D(x)] + E_{z\sim P_z(z)}[1-\log D(G(z))]$$

记 $V(D,G) = E_{x\sim P_{data}}[\log D(x)] + E_{x\sim P_g}[1-\log D(x)]$

① for fixed G, 求 $\max_D V(D,G)$

$$\max_D V(D,G) = \int P_{data}\log D(x)\,dx + \int P_g(1-\log D(x))\,dx$$

$$= \int [P_d \log D(x) + P_g(1-\log D(x))]\,dx$$

$$\frac{\partial}{\partial D} = \int \frac{\partial}{\partial D}[P_d \log D(x) + P_g(1-\log D(x))]\,dx$$

$$= \int \left[P_d \frac{1}{D} + P_g \cdot \frac{-1}{1-D}\right] dx = 0$$

$$\Rightarrow D_G^* = \frac{P_d}{P_d + P_g}$$

② 将 D_G^* 代入 $V(D,G)$

$$\min_G \max_D V(D,G) = \min_G V(D_G^*, G)$$

$$\min_G V(D_G^*, G)$$

$$= \min_G E_{x\sim P_d}\left[\log \frac{P_d}{P_d+P_g}\right] + E_{x\sim P_g}\left[\log \frac{P_g}{P_d+P_g}\right]$$

$$= \min_G E_{x\sim P_d}\left[\log \left(\frac{P_d}{(P_d+P_g)/2}\cdot\frac{1}{2}\right)\right] +$$

$$E_{x\sim P_g}\left[\log \left(\frac{P_g}{(P_d+P_g)/2}\cdot\frac{1}{2}\right)\right]$$

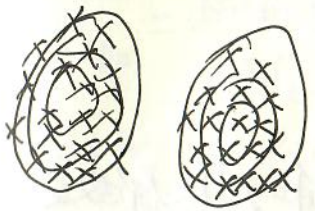
$$= \min_G KL\left(P_d \,\|\, \frac{P_d+P_g}{2}\right) + KL\left(P_g \,\|\, \frac{P_d+P_g}{2}\right) - \log 4$$

$$\geq \log 4$$

当 $P_d = \frac{P_d+P_g}{2} = P_g$ 时，等号成立

$$P_g^* = P_d,\quad D_G^* = \frac{P_d}{P_d+P_g} = \frac{1}{2}.$$

(二十二) 谱聚类 Spectral Clustering.



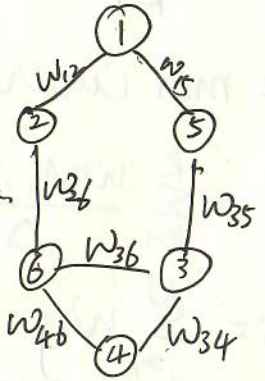
⇒ compactness:
K-means GMM.

⇒ Connectivity:
spectral clustering.

Graph-based (带权重的无向图)

$G = \{V, E\}$

$V = \{1, 2, \cdots, N\}$ (顶点集合)
一个节点一个样本



E: $W = [w_{ij}]$, $1 \leq i,j \leq N$
(N个顶点)

W: Similarity matrix (affinity matrix)

$$w_{ij} = \begin{cases} K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), & (i,j) \in E \\ \end{cases}$$

定义：
$A \subset V$, $B \subset V$, $A \cap B = \emptyset$

$$W(A,B) = \sum_{\substack{i \in A \\ j \in B}} w_{ij}$$

（$W(A,B)$是有相连的节点的和，可以看做是将集合切割以后的损失）

假如一共K个类别

$cut(V) = cut(A_1, A_2, \cdots, A_K)$ （将V分成K个子集）

$$V = \bigcup_{k=1}^{K} A_k \quad A_i \cap A_j = \emptyset, \ \forall i,j \in \{1,2,\cdots,K\}$$

$$cut(V) = \sum_{k=1}^{K} W(A_k, \bar{A}_k) = \sum_{k=1}^{K} W(A_k, V) - W(A_k, A_k)$$

目标：$\min Cut(V) \ \{A_k\}_{k=1}^{K}$

$$Ncut = \sum_{k=1}^{K} \frac{W(A_k, \bar{A}_k)}{\Delta} \qquad \Delta = degree(A_k) = \sum_{i \in A_k} d_i$$

$$d_i = \sum_{j=1}^{N} w_{ij}$$

$$Ncut = \sum_{k=1}^{K} \frac{W(A_k, \bar{A}_k)}{\sum_{i \in A_k} d_i}$$

Model:

$$\{\hat{A}_k\}_{k=1}^{K} = \underset{\{A_k\}_{k=1}^{K}}{\arg\min} \ Ncut(V)$$

indicator vector:

$$\begin{cases} y_i \in \{0,1\}^K & y_{ij} = 1 \Leftrightarrow \text{第} i \text{个样本属于} j \text{个类别} \\ \sum_{j=1}^{K} y_{ij} = 1 & 1 \le i \le N \\ & 1 \le j \le K \end{cases}$$

$$Y = (y_1, y_2, \cdots, y_N)^T_{N \times K} \qquad y_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \Big\} K$$

$$\hat{Y} = \underset{Y}{\arg\min} \ Ncut(V)$$

$$Ncut(V) = \sum_{k=1}^{K} \frac{W(A_k, \bar{A}_k)}{\sum_{i \in A_k} d_i}$$

$$= tr\begin{pmatrix} \frac{W(A_1, \bar{A}_1)}{\sum_{i \in A_1} d_i} & & 0 \\ & \ddots & \\ 0 & & \frac{W(A_K, \bar{A}_K)}{\sum_{i \in A_K} d_i} \end{pmatrix}$$

$$= tr\begin{pmatrix} W(A_1, \bar{A}_1) & & \\ & \ddots & \\ & & W(A_K, \bar{A}_K) \end{pmatrix} \cdot \begin{pmatrix} \sum_{i \in A_1} d_i & & \\ & \ddots & \\ & & \sum_{i \in A_K} d_i \end{pmatrix}^{-1}$$

已知 W, Y, 求 D, P.

$$\underbrace{Y^TY}_{K\times K} = (y_1, y_2, \cdots, y_N)\begin{pmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_N^T \end{pmatrix} = \sum_{i=1}^{N} y_i y_i^T$$

$$= \begin{pmatrix} N_1 & & \\ & N_2 & \\ & & \ddots \\ & & & N_K \end{pmatrix}_{K\times K} = \begin{pmatrix} \sum_{i\in A_1} 1 & & \\ & \ddots & \\ & & \sum_{i\in A_K} 1 \end{pmatrix}_{K\times K}$$

N_k: 在 N 个样本中, 属于 K 的样本个数.

$$\sum_{k=1}^{K} N_k = N, \quad N_k = |A_k| = \sum_{i\in A_k} 1$$

$$\sum_{i=1}^{N} y_i y_i^T d_i = \sum_{i=1}^{N} y_i d_i y_i^T$$

$$P = \begin{pmatrix} \sum_{i\in A_1} d_i & & \\ & \ddots & \\ & & \sum_{i\in A_K} d_i \end{pmatrix} \qquad D = \begin{pmatrix} d_1 & & \\ & d_2 & \\ & & \ddots \\ & & & d_N \end{pmatrix}$$

$$= Y^T \cdot D \cdot Y \qquad = \mathrm{diag}(W \cdot 1_N)$$

$$O = \begin{pmatrix} w(A_1, \bar{A}_1) & & \\ & \ddots & \\ & & w(A_K, \bar{A}_K) \end{pmatrix}$$

$\sum_{i\in A_1} d_i$ → $w(A_1, \bar{A}_1) = w(A_1, V) - w(A_1, A_1)$ ← $\sum_{i\in A_1}\sum_{j\in A_1} w_{ij}$

$$O = \begin{pmatrix} \sum_{i\in A_1} d_i & & \\ & \ddots & \\ & & \sum_{i\in A_K} d_i \end{pmatrix} - \begin{pmatrix} w(A_1, A_1) & & \\ & \ddots & \\ & & w(A_K, A_K) \end{pmatrix}$$

$$0 = Y^TDY - Y^TWY$$

$$Y^TWY = (y_1 \cdots y_N)\begin{pmatrix} w_{11} & \cdots & w_{1N} \\ \vdots & & \vdots \\ w_{N1} & \cdots & w_{NN} \end{pmatrix}\begin{pmatrix} y_1^T \\ \vdots \\ y_N^T \end{pmatrix}$$

$$= \left(\sum_{i=1}^{N} y_i w_{i1} \cdots \sum_{i=1}^{N} y_i w_{iN}\right)\begin{pmatrix} y_1^T \\ \vdots \\ y_N^T \end{pmatrix}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j^T w_{ij}$$

$$= \begin{pmatrix} \sum_{i\in A_1}\sum_{j\in A_1} w_{ij} & \cdots & \sum_{i\in A_1}\sum_{j\in A_K} w_{ij} \\ \vdots & & \vdots \\ \sum_{i\in A_K}\sum_{j\in A_1} w_{ij} & \cdots & \sum_{i\in A_K}\sum_{j\in A_K} w_{ij} \end{pmatrix}$$

$$Ncut(V) = tr(O \cdot P)$$
$$= tr(O' \cdot P)$$

$$\hat{Y} = \arg\min_{Y} tr\left(Y^T(D-W)Y \cdot (Y^T D Y)^{-1}\right)$$

$L = D - W$ (Laplacian Matrix).

(三十二) VAE

Variational Auto-encoder.

Latent Variable Model

$Z \rightarrow X$

GMM: K个Gaussian Dist混合: $Z \sim$ Categorical Dist

VAE: infinite Gaussian Dist混合: $Z \sim N(0, I)$ (解空间高斯分布, 单位矩阵)

$X|Z \sim N(\mu_\theta(Z), \Sigma_\theta(Z))$

$$P(X) = \int_Z P(X, Z)\, dZ = \int_Z P(Z) \cdot P(X|Z)\, dZ$$

↓ intractable

$$P(Z|X) = \frac{P(Z) \cdot P(X|Z)}{P(X)}$$

↓ intractable

$P_\theta(X|Z)$

$P(Z) = N(0, I)$

$P_\theta(X|Z) = N(\mu_\theta(Z), \Sigma_\theta(Z))$

$P_\theta(Z|X)$ is intractable.

$q_\phi(Z|X)$

EM: $\log P(X) = ELBO + KL(q_\phi(Z|X) \,||\, P_\theta(Z|X))$

E-step: 当 $q = P_\theta(Z|X)$ $KL = 0$, $E = ELBO$

M-step: $\theta = \arg\max ELBO$

$$\langle \hat{\theta}, \hat{\phi} \rangle = \arg\max KL(q_\phi(z|x) \| p_\theta(z|x))$$

$$= \arg\max ELBO$$

$$= \arg\max E_{q_\phi(z|x)}[\log p_\theta(x,z)] + H[q_\phi]$$

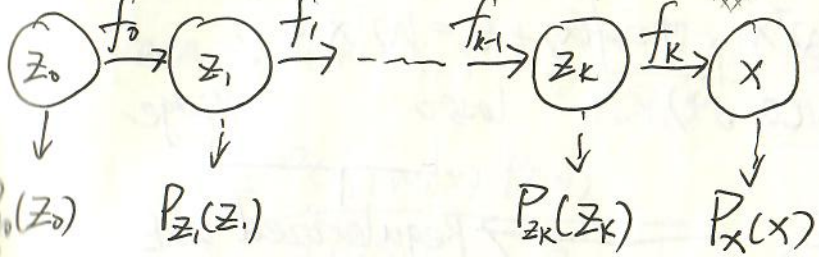
$$= \arg\max E_{q_\phi(z|x)}[\log p_\theta(x|z) + \log p_\theta(z)] + H[q_\phi]$$

$$= \arg\max E_{q_\phi(z|x)}[\log p_\theta(x|z)] + E_{q_\phi(z|x)} \log p_\theta(z) + E_{q_\phi(z|x)} \log q_\phi(z|x)$$

$$= \arg\max E_q[\log p_\theta(x|z)] - KL(q_\phi(z|x) \| p(z))$$

$\hat{\phi}$ 熵最大.

(三十三) Normalizing Flow 流模型



change of variable theorem.

$x = f(z)$, $z, x \in \mathbb{R}^p$

$z \sim p_z(z)$, $x \sim p_x(x)$

f is continuous, invertible.

$\therefore \int_z p_z(z)\,dz = 1 = \int_x p_x(x)\,dx \Rightarrow |p_z(z)\,dz| = |p_x(x)\,dx|$

$\therefore p_x(x) = \left|\frac{dz}{dx}\right| \cdot p_z(z)$

$\because x = f(z)$, f is invertible

$\therefore z = f^{-1}(x)$

$\therefore p_x(x) = \left|\frac{\partial f^{-1}(x)}{\partial x}\right| \cdot p_z(z)$ $\frac{\partial f^{-1}(x)}{\partial x}$ Jacobian Matrix

$p_x(x) = \left|\det\left(\frac{\partial f^{-1}(x)}{\partial x}\right)\right| \cdot p_z(z)$ $\det\left(\frac{\partial f^{-1}(x)}{\partial x}\right)$ Jacobian Det

$= \left|\det\left(\frac{\partial f(z)}{\partial z}\right)^{-1}\right| \cdot p_z(z)$

(十九) 贝叶斯线性回归

$f(x) = w^T x$, $y = f(x) + \varepsilon = w^T x + \varepsilon$

$\varepsilon \sim N(0, \sigma^2)$

LSE (最小二乘) → Regularized LSE (Lasso, Ridge)

LSE ↑ MLE (noise is Gaussian)

Regularized LSE ↓ MAP (noise is Gaussian)

$$w_{MLE} = \arg\max_w P(Data|w)$$

$$P(w|Data) \propto P(Data|w) P(w)$$

$$w_{MAP} = \arg\max_w P(w|Data)$$

$$= \arg\max_w P(Data|w) P(w)$$

$P(w)$ → $N(\mu_0, \Sigma_0)$

Inference:

$$P(w|Data) = P(w|X, Y) = \frac{P(w, Y|X)}{P(Y|X)} = \frac{P(w, Y|X)}{P(Y|w, X) P(w|X)} / P(Y|X)$$

$$= \frac{P(Y|w, X) P(w)}{\int P(Y|w, X) \cdot P(w) \, dw}$$

$P(w) = P(w|X)$ 因为 w 定好了分布.

$$P(Y|w, X) = \prod_{i=1}^{N} P(y_i|w, x_i) = \prod_{i=1}^{N} N(y_i|w^T x_i, \sigma^2)$$

$P(y|x, w) \sim N(w^T x, \sigma^2)$

$P(w) = N(0, \Sigma_p)$

$P(w|Data) \propto P(Y|w, X) \cdot P(w)$

Gaussian

$$\propto \prod_{i=1}^{N} N(y_i|w^T x_i, \sigma^2) \cdot N(0, \Sigma_p)$$

共轭: Gaussian 分布自共轭

Softmax Classifier Vectorized.

$$L_i = -\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right) = -f_{y_i} + \log\sum_j e^{f_j}$$

$f_j = x_i w_j$ $\quad x_i w$

$$\frac{dL_i}{dw_j} = \frac{1}{\sum_j e^{f_j}} \cdot x_i \cdot e^{x_i w_j} = \boxed{\frac{e^{f_j}}{\sum_j e^{f_j}} \cdot x_i}$$

$$\frac{dL_i}{dw_{y_i}} = -x_i + \frac{e^{f_{y_i}}}{\sum_j e^{f_j}} \cdot x_i = \left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}} - 1\right) x_i$$

$$\frac{dL_i}{dw} = \begin{bmatrix} \frac{dL_i}{dw_1} & \cdots & \frac{dL_i}{dw_j} \\ \vdots & & \vdots \\ \frac{dL_i}{dw_1} & \cdots & \frac{dL_i}{dw_j} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{e^{x_1 w_1}}{\sum_j e^{x_1 w_j}} \cdot x_1 & \cdots & \frac{e^{x_1 w_j}}{\sum_j e^{x_1 w_j}} \cdot x_1 \\ \frac{e^{x_2 w_1}}{\sum_j e^{x_2 w_j}} \cdot x_2 & \cdots & \frac{e^{x_2 w_j}}{\sum_j e^{x_2 w_j}} \cdot x_2 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{e^{x_1 w_1}}{\sum e^{x_1 w_j}} & \cdots & \frac{e^{x_1 w_j}}{\sum e^{x_1 w_j}} \\ \vdots & & \vdots \\ \frac{e^{x_2 w_1}}{\sum e^{x_2 w_j}} & \cdots & \frac{e^{x_2 w_j}}{\sum e^{x_2 w_j}} \end{bmatrix} \cdot \begin{bmatrix} x_1 & & \\ & \ddots & \\ & & x_2 \end{bmatrix}$$

$$= [x_1, x_2 \cdots, x_i] \cdot \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}$$

$$\frac{dL_i}{dw} = \left[\frac{e^{f_j}}{\sum e^{f_j}} \frac{dL_i}{dw_1} \cdots \frac{dL_i}{dw_j}\right]$$

$$= \left[\frac{e^{x_i w_1}}{\sum_j e^{f_j}} \cdot x_i \cdots \right]$$

$$= \left[\frac{e^{x_i w_1}}{\sum e^{f_j}} \cdots \frac{e^{x_i w_j}}{\sum e^{f_j}}\right] \cdot [x_i \cdots x_i]$$