机器学习笔记

Yuxuan Yuan & Lulu Cao

2022.2.21

目录

1 Lecture 1

2022.2.21 第一节课是在 C103 上的，啥也没带 oh lambda 含义

2 Lecture 2

2022.2.28 yyx 忘了记录板书 oh

2.1 课堂回顾

机器学习 =lambda ：机器 = 函数；学习 = 拟合

Machine Learning = LAMBDA. Loss,Algorithm, Model, BigData, Application。

BigData $D=\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$，其中 $x_n \in R^N$，$y \in R$

$$\boldsymbol{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \boldsymbol{w}=\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}, \quad \boldsymbol{X}=\begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_N^{\mathrm{T}} \end{bmatrix}$$

Model

$$y=\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}$$

Loss

$$\mathcal{L}_2(\boldsymbol{w})=\frac{1}{N}\sum_{n=1}^{N}\left(y_n-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n\right)^2$$

将上述式子矩阵化，

$$\begin{aligned}\mathcal{L}_2(\boldsymbol{w}) &= \frac{1}{N}(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}) \\ &= \frac{1}{N}\left((\boldsymbol{X}\boldsymbol{w})^{\mathrm{T}}-\boldsymbol{y}^{\mathrm{T}}\right)(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}) \\ &= \frac{1}{N}(\boldsymbol{X}\boldsymbol{w})^{\mathrm{T}}\boldsymbol{X}\boldsymbol{w}-\frac{1}{N}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{w}-\frac{1}{N}(\boldsymbol{X}\boldsymbol{w})^{\mathrm{T}}\boldsymbol{y}+\frac{1}{N}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} \\ &= \frac{1}{N}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{w}-\frac{2}{N}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}+\frac{1}{N}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}\end{aligned}$$

Algorithm 对损失函数求导，并令其等于 0

$$\begin{aligned}\frac{\partial \mathcal{L}_2}{\partial w} &= \frac{2}{N}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{w}-\frac{2}{N}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}=0 \\ \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{w} &= \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}\end{aligned}$$

矩阵求导相关

$f(w)$	$\frac{\partial f}{\partial w}$
$w^{\mathrm{T}}\boldsymbol{x}$	$\boldsymbol{x}$
$\boldsymbol{x}^{\mathrm{T}}w$	$\boldsymbol{x}$
$w^{\mathrm{T}}w$	$2w$
$w^{\mathrm{T}}\boldsymbol{C}w$	$2\boldsymbol{C}w$

从而，

$$\boldsymbol{w}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

2.2 投影矩阵 & Normal Equation

2.3 增加新的数据 (样本数 or 特征维度)

作为作业，当 $\boldsymbol{X}$ 增加一行 (样本数增加) 或者增加一列 (特征维度增加) 时，$\boldsymbol{W}$ 如何变化，写出更新后的 $\boldsymbol{W}$ 和更新前的 $\boldsymbol{W}$ 之间的增量表达式。

相关公式

- **Sherman-Morrison 公式**
$$\left(A+uv^T\right)^{-1}=A^{-1}-\frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}$$
- **分块矩阵求逆** 设 A 是 $m\times m$ 可逆矩阵, B 是 $m\times n$ 矩阵, C 是 $n\times m$ 矩阵, D 是 $n\times n$ 矩阵, $D-CA^{-1}B$ 是 $n\times n$ 可逆矩阵, 则有
$$\left[\begin{array}{ll} A & B \\ C & D \end{array}\right]^{-1}=\left[\begin{array}{cc} A^{-1}+A^{-1}B\left(D-CA^{-1}B\right)^{-1}CA^{-1} & -A^{-1}B\left(D-CA^{-1}B\right)^{-1} \\ -\left(D-CA^{-1}B\right)^{-1}CA^{-1} & \left(D-CA^{-1}B\right)^{-1} \end{array}\right]$$

2.4 人工智能的流派

- 类比主义：核方法、SVM
- 连接主义：神经网络
- 贝叶斯主义
- 符号主义：决策树、专家系统
- 演化主义（优化算法）：遗传算法
- 行为主义：强化学习

2.5 人脸识别

设 D 为人脸的数据, $D=\{(\boldsymbol{x}_n,y_n)\}_{n=1}^N$，其中 $x_n\in R^N$，$y\in[1,100]$, $y_i=w^Tx_i+\epsilon,\epsilon\sim W(0,\sigma^2)$，$\sigma^2=1$, 则 $y_i\sim W(w^Tx_i,1)$

方差一样，所以一样胖；均值不一样，所以 location 不同

找到一个 w，使得 y_i 出现的概率最大, 即 $P(y_i|w^Tx,1)=\frac{1}{\sqrt{2\pi}}exp-\frac{1}{2}(\frac{y_i-w^Tx}{1})^2$

$$L=p\left(\boldsymbol{y}\mid\boldsymbol{X},w,\sigma^2\right)=\prod_{n=1}^N p\left(y_n\mid\boldsymbol{x}_n,w,\sigma^2\right)=\prod_{n=1}^N\mathcal{N}\left(w^{\mathrm{T}}\boldsymbol{x}_n,\sigma^2\right)$$

取对数, 有

$$\begin{aligned}\log L&=\sum_{n=1}^N\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}\left(y_n-f\left(\boldsymbol{x}_n;\boldsymbol{w}\right)\right)^2\right\}\right)\\&=\sum_{n=1}^N\left(-\frac{1}{2}\log(2\pi)-\log\sigma-\frac{1}{2\sigma^2}\left(y_n-f\left(\boldsymbol{x}_n;w\right)\right)^2\right)\\&=-\frac{N}{2}\log 2\pi-N\log\sigma-\frac{1}{2\sigma^2}\sum_{n=1}^N\left(y_n-f\left(\boldsymbol{x}_n;w\right)\right)^2\end{aligned}$$

$w^{\mathrm{T}}\boldsymbol{x}_n$ 替换模型中的决定性部分, 对数似然表达式就呈现如下的形式: $\log L = -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(y_n - w^{\mathrm{T}}\boldsymbol{x}_n\right)^2$ 得到的最小二乘解, 通过求导数、使其等于零以及求解拐点的方法, 类似于 1.1.4 节所述的方式。对于 w (注意, $w^{\mathrm{T}}\boldsymbol{x}_n = \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{w}$),

最小二乘法和极大似然估计是等价的

3 Lecture 3

2022.3.7

Outline

1. Least Square: MLE

    1.1 SGD

    1.2 Probalistic Graph Representation

    1.3 $\mathbb{E}[\hat{w}]/cov[\hat{w}]$

    1.4 bias / variance

2. Revisted LS: Curve Fitting

    2.1 Model Selection

    2.2 Overfitting

3. How to solve overfitting

    3.1 Regularization (MAP)

    3.2 Bayesian Learning

4 Lecture 4

2022.3.14

使用 MLE/MAP/Bayes Learning 三种方法来求解参数，通过 4 个例子来加深。Example 4 没讲完。其中 Example 3 的三种解法需要自己课后补充。

tips: to be added 三个图对应生成式模型、判别式模型、分布计算; 一些前置 discrete continuous 的概率表变量分布；有积分的地方和是求谁的期望的地方，

Outline

- MLE / MAP / Bayesian
- Example 1. Bernoulli
- Example 2. Gaussian
- Example 3. Linear Regression
- Example 4. Logistic Regression

问题:

- 共轭分布是什么意思
- β 分布

4.1 Example 1. Bernoulli

Given dataset $D = \{< x_i >\}_{i=1}^{n}, x_i \in \{0, 1\}$ x_i 服从 Bernoulli 分布，$P(x_i = 1) = \theta$; 再假设 n_1 表示 D 中 $x_i = 1$ 的个数；

Method

1. MLE

$$
\begin{aligned}
P(D|\theta) &= \prod_{i=1}^{n} P(x_i;\theta) \\
&= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\
&= \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i} \\
&= \theta^{n_1}(1-\theta)^{n-n_1}
\end{aligned}
$$

取对数以便计算，从而

$$
\arg\max_{\theta} \ln P(D|\theta) = \arg\max_{\theta} n_1 \ln\theta + (n-n_1)\ln(1-\theta)
$$

对 θ 求导

$$
\frac{n_1}{\theta} - \frac{n-n_1}{1-\theta} = 0
$$

得到 $\theta = \frac{n_1}{n}$

2. MAP

Beta 分布 $Beta(\theta \mid a, b)$ $P(\theta|D) \propto P()$

贝塔分布（Beta Distribution) 是一个作为伯努利分布和二项式分布的共轭先验分布的密度函数。在概率论中，贝塔分布，也称 B 分布，是指一组定义在 (0,1) 区间的连续概率分布。

3. Bayesian

4.2 Example 2. Gaussian

Method

1. MLE
2. MAP
3. Bayesian

4.3 Example 3. Linear Regression

Given dataset $D = \{< x_i, y_i >\}_{i=1}^{n}, y_i \in \mathbb{R}$, 假设 $X\ \mathcal{N}(\theta, \sigma^2)$，则有 $P(x;\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{(x-\theta)^2}{2\sigma^2}\}$

Method

1. MLE
2. MAP
3. Bayesian

4.4 Example 4. Logistic Regression

Method

1. MLE

2. MAP 3. Bayesian

计算步骤 (流程) 总结

1. $P(D|\theta)$
2. $P(\theta)$
3. $P(\theta|D) \propto P(D|\theta)P(\theta)$
4. $\theta_{bayes} = \mathbb{E}[\theta|D]$
5. inference: $P(y_{new}|D, x_{new}) = \int p(y_{new}|x_{new}, \theta)p(\theta|D)$

其中 $D = (X, Y)$

5 Lecture 5

2022.3.21

Outline

1. Review: MLE / MAP / Bayesian Estimation
2. Logistic Regression
3. Perceptron
4. Generative Classification Model

   x is continuous (GDA)

   x is discrete (NB)

....

Data: $D = < x_i, y_i >_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in 0, 1, D = (X, Y)$

Model: $P(x, y; \Theta)$(生成式)；$P(y|x; \Theta)$(判别式)；$P(x_i; \Theta)$ 无监督

Inference: Given x, D; Output y, $P(y|x, D) =?$

Learning: Given D; Output Θ, $P(D|\Theta)$, $\Theta_{Bayes} = \mathbb{E}[\Theta|D]$

....

5.1 Review: MLE / MAP / Bayesian

三种方法，即 Learning 的三种方法

Learning

1. MLE

$$\begin{aligned}\hat{\Theta}_{MLE} &= \arg\max_{\Theta} P(D|\Theta) \\ &= \arg\max_{\Theta} \sum_{i=1}^{n} \ln P(blank; \Theta)\end{aligned}$$

其中，$blank$ 可以填入 1)x_i, y_i; 2)$y_i|x_i$; 3)x_i; 即三种模型都可用 MLE 求解

2. MAP

$$\begin{aligned}\hat{\Theta}_{MAP} &= \arg\max_{\Theta} P(\Theta|D) \propto P(D|\Theta)P(\Theta) \\ &= \arg\max_{\Theta} \ln P(blank; \Theta) + \ln P(\Theta)\end{aligned}$$

其中，$P(\Theta)$ 是先验，$P(D|\Theta)$ 是似然，等式的最后 $\ln P(blank$ 为数据项（Data Term），$\ln P(\Theta)$ 为正则项 (Regularization Term, 或平滑项 (Smooth Term))。

* 当 $P(\Theta)$ 是均匀分布时，$MLE = MAP$ 。

3. Bayesian Estimation

(前提是 $P(\Theta|D)$ 即后验分布已知)

$$\hat{\Theta}_{Bayes} = \mathbb{E}[\Theta|D] = E_{\theta \sim P(\cdot|D)}[\Theta] = \int \Theta p(\Theta|D) d\Theta$$

Inference

利用上述三种方法做参数估计后的推理方法

1. MLE

Given $\hat{\Theta}_{MLE}, x, D$, ouput $P(y|x;\hat{\Theta}_{MLE})$.

举例，在 Logistic Regression 中，$P(y=1|x;\hat{\Theta}_{MLE}) = \sigma(\hat{\Theta}_{MLE}^T x)$

2. MAP

Given $\hat{\Theta}_{MAP}, x, D$, output $P(y|x;\hat{\Theta}_{MAP})$.

举例，在 Logistic Regression 中，$P(y=1|x;\hat{\Theta}_{MAP}) = \sigma(\hat{\Theta}_{MAP}^T x)$

3.Bayes Estimation

Given $x, D, P(\Theta|D)$, ouput $P(y|x;\hat{\Theta}_{MAP})$.

$$
\begin{aligned}
P(y|x;D) &= \int p(y,\Theta|x;D)d\Theta \\
&= \int p(\Theta|x;D)p(y|\Theta,x;D)d\Theta
\end{aligned}
$$

等式最后的 $p(\Theta|x;D)$ 为后验分布，$p(y|\Theta,x;D)$ 为模型。能这么做的前提是假设 x_{new} 与 Θ 无关。

如果求 $y=1$, 有

$$
\begin{aligned}
p(y|\Theta,x;D) &= \int p(\Theta|D)\sigma(\Theta^T x)dx (= \mathbb{E}_{\Theta\sim P(\cdot|D)}[\sigma(\Theta^T x)] \\
&\approx \frac{1}{\mathcal{L}}\sum_{i=1}^{\mathcal{L}} \sigma((\Theta^{(i)})^T x)
\end{aligned}
$$

其中有假设 x_{new} 与 Θ 相互独立；$\Theta^{(i)} \sim P(\Theta|D))), i = 1, ..., \mathcal{L}$, 上面公式使用了抽样技术 (Sampling) 来求期望。

5.2 Logistic Regression(逻辑斯蒂回归)

补点图概率图模型（可观测量用灰色阴影）

Learning

1.MLE

$$P(y_i|x_i;\Theta) = \sigma(\Theta^T x_i)^{y_i}(1-\sigma(\Theta^T x_i))^{1-y_i}$$

$$
\begin{aligned}
lnP(D|\Theta) &= \sum_{i=1}^{n}\{y_i \ln\sigma_i + (1-y_i)\ln(1-\sigma_i)\} \\
&= \mathcal{L}_D(\Theta)
\end{aligned}
$$

其中 $\sigma_i = \sigma(a_i) = \sigma(\Theta^T x_i) = \frac{1}{1+e^{\Theta^T x_i}}$，从而

$$\Theta^* = \arg\max_{\Theta} \mathcal{L}_D(\Theta)$$

到这里，求解 Θ^* 方法有求偏导数并等于 0 来计算解析解，但无法做到，原因如下, 利用链式法则算一下偏导数

$$\begin{aligned}\nabla_\Theta \mathcal{L}_D(\Theta) &= \frac{\partial \mathcal{L}_D(\Theta)}{\partial \Theta} \\ &= \frac{\partial \mathcal{L}_D(\Theta)}{\partial \sigma_i}\frac{\partial \sigma_i}{\partial a_i}\frac{\partial a_i}{\partial \Theta} \\ &= \sum_{i=1}^{n}(\frac{y_i}{\sigma_i} - \frac{1-y_i}{1-\sigma_i})\sigma_i(1-\sigma_i)x_i \\ &= \sum_{i=1}^{n}(y_i - \sigma_i)x_i\end{aligned}$$

试试，几乎无法求得解析解。

第二种方法，尝试二阶导数 $\nabla_\Theta(\nabla_\Theta \mathcal{L}_D(\Theta))$，令 $\nabla_\Theta \mathcal{L}_D(\Theta) = g$,

Algorithm 1 A1: Gradient Ascent for Logistic Regression

Input: X, Y;

Output: Θ;

init: $\Theta \sim \mathcal{N}(0, \alpha^{-1}\mathrm{I}), \epsilon$

Loop:

$g = \mathbf{X}^T(\mathbf{Y} - \sigma)$

$\Theta^{t+1} := \Theta^t + \eta g$

until $\| \Theta^{t+1} - \Theta^t \| \leq \epsilon$

return Θ;

Hessian Matrix of the Loss Function $\mathcal{L}_D(\Theta)$

Newton $\Theta^{t+1} := \Theta^t + \mathrm{H}^{-1}g$ ......todo

2.MAP

$$\begin{aligned}P(\Theta) &\sim \mathcal{N}(0, \alpha^{-1}\mathrm{I}) \\ P(D|\Theta) &= \prod_{i=1}^{n}\sigma_i^{y_i}(1-\sigma_i)^{1-y_i} \\ P(\Theta|D) &\propto \mathcal{N}(0, \alpha^{-1}\mathrm{I})\prod_{i=1}^{n}\sigma_i^{y_i}(1-\sigma_i)^{1-y_i}\end{aligned}$$

根据 MAP，有

$$\begin{aligned}\Theta^* &= \arg\max_\Theta \ln P(\Theta|D) \\ &= \arg\max_\Theta \ln P(D|\Theta) + \ln P(\Theta) \\ &= \arg\max_\Theta \sum_{i=1}^{n} y_i \ln(\sigma_i) + (1-y_i)\ln(1-\sigma_i) - \frac{1}{2}(\Theta^{-1}\Sigma^{-1}\Theta)\end{aligned}$$

则

$$
\begin{aligned}
\nabla_\Theta \mathcal{L}_D(\Theta) &= \frac{\partial \mathcal{L}_D(\Theta)}{\partial \Theta} \\
&= \sum_{i=1}^{n} (y_i - \sigma_i) x_i - \Theta \Sigma^{-1}
\end{aligned}
$$

这里定 $\alpha^{-1}\mathrm{I} = \Sigma$, 不影响结果。

类似算法A1, 写个A2

Algorithm 2 A2: (MAP) Gradient Ascent for Logistic Regression

Input: X, Y;

Output: Θ;

1: init: $\Theta \sim \mathcal{N}(0, \alpha^{-1}\mathrm{I}), \epsilon$
2: **Loop:**
3: $\quad g = \mathbf{X}^T(\mathbf{Y} - \sigma) - \Theta^t \Sigma^{-1}$
4: $\quad \Theta^{t+1} := \Theta^t + \eta g$
5: **until** $\|\ \Theta^{t+1} - \Theta^t\ \| \leq \epsilon$
6: **return** Θ;

3.Bayesian Estimation

贝叶斯估计：$\mathbb{E}[\Theta|D]$

$P(y=1|x, D) =$ 类似算法A1, 写个A3用于贝叶斯推理

Inference 后验预测分布为

$$
p(y=1|x_{new}, D) = \int p(y=1|x_{new}, \Theta) p(\Theta|x, D) d\Theta
$$

但是积分难以处理，采用近似处理，同时还有假设 x_{new} 与 Θ 无关，有

$$
\begin{aligned}
p(y=1|x_{new}, D) &= \int p(y=1|x_{new}, \Theta, D) p(\Theta|D) d\Theta \\
&= \int \sigma(\Theta^T x_{new}) p(\Theta|D) d\Theta \\
&\approx \frac{1}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \sigma((\Theta^{(i)})^T x_{new})
\end{aligned}
$$

其中，$\Theta^{(i)} \propto p(\Theta|D)(i = 1, ..., \mathcal{L})$

Algorithm 3 A3: Inference after Bayesian Estimation

Input: x_{new}, D;

Output: y;

1: init: $\Theta^{(i)} \propto p(\Theta|D)(i = 1, ..., \mathcal{L})$
2: $P(y=1|x_{new}, D) = \frac{1}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \sigma((\Theta^{(i)})^T x_{new})$
3: $P(y=0|x_{new}, D) = \frac{1}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} (1 - \sigma((\Theta^{(i)})^T x_{new}))$
4: **return** 1 if $P(y=1|x_{new}, D) > P(y=0|x_{new}, D)$ else 0;

5.3 Perceptron(感知机)

写个历史表

- 1943 M-P —-神经元模型
- 1957 Rosenblatt —-Perceptron;
- 1982 BP 算法 (Computational Graph 计算图)
- 2006 Hinton —-预训练 pretain、DNN
- 2012 AlexNet.... 图像不太懂...

Model
计算方法：$y = sgn(\mathbf{W}^T x + b)$
其中;

$$\begin{cases} y = +1, \mathbf{W}^T x + b > 0, \\ y = -1, \mathbf{W}^T x + b < 0 \end{cases}$$

Loss 计算

$$\mathcal{L}_D(W) = \sum_{x_i, y_i \in M} \frac{|y_i(W^T x_i + b)|}{||W||}$$

其中 $M = \{(x_i, y_i) | y_i(W^T x + b) < 0\}$，可令 $||W|| = 1$，上式可写成,

$$\mathcal{L}_D(W) = \sum_{i=1}^{n} \max\{0, -y_i(W^T x_i + b)\}$$

* 当 max 中的第二项变为 $1 - y_i(W^T x_i + b)$ 时，损失函数变成 SVM 的损失函数。

这里求一下偏导数, 对于单个样本 (x_i, y_i)

$$\frac{\partial(-y_i(W^T x_i + b))}{\partial W} = -y_i x_i$$

也可以写成一个算法形式

Algorithm 4 A4: Perceptron GD

Input: X, Y;
Output: $\Theta = \{W, b\}$;

```
init: W, b
Loop:
  if y_i(W^T x_i + b) < 0 then
    W^{t+1} ← W^t − η∇_W L_D(W)
    等价于 W^{t+1} ← W^t + η y_i x_i
  end if
until 训练集中没有误分类点
return Θ;
```

5.4 Generative Classification Model

y 是离散的

x 是连续的

Given $\{(x_i, y_i)\}_{i=1}^n = D = (X, Y), x \in \mathbb{R}^d$。假设 $y_i = 0, 1$，即 y_i 服从 Bernoulli 分布，x_i 服从高斯分布 $x_i|y_i \sim \mathcal{N}(\mu_k, \Sigma_k)$，model 描述如下

$$\begin{aligned} P(x_i, y_i; \Theta) &= P(y_i; \Theta)P(x_i|y_i; \Theta) \\ &= \text{Bernoulli}(y_i|p) \prod_{y=0}^{1} \mathcal{N}(x_i|\mu_k, \Sigma_k)^{\mathbf{1}\{k=y\}} \end{aligned}$$

关于 $P(x_i|y_i; \Theta)$ 部分细述如下,

$$\begin{aligned} P(x|y = 1; \Theta) &= \mathcal{N}(x|\mu_1, \Sigma_1) \\ P(x|y = 0; \Theta) &= \mathcal{N}(x|\mu_0, \Sigma_0) \end{aligned}$$

从而有

$$P(x|y; \Theta) = \mathcal{N}(x|\mu_1, \Sigma_1)^{\mathbf{1}\{y=1\}} \mathcal{N}(x|\mu_0, \Sigma_0)^{\mathbf{1}\{y=0\}}$$

记 $\Theta = (p, \mu_0, \Sigma_0, \mu_1, \Sigma_1)$，则有

$$\begin{aligned} P(D; \Theta) &= \prod_{i=1}^{n} P(x_i, y_i; \Theta) \\ &= \prod_{i=1}^{n} (p^{y_i}(1-p)^{1-y_i}) \mathcal{N}(x|\mu_1, \Sigma_1)^{\mathbf{1}\{y=1\}} \mathcal{N}(x|\mu_0, \Sigma_0)^{\mathbf{1}\{y=0\}} \end{aligned}$$

取对数，有

$$\ln P(D; \Theta) = \sum_{i=1}^{n} y_i \ln p + (1 - y_i) \ln(1 - p) + \sum_{i=1}^{n} \mathbf{1}\{y = 1\} \ln \mathcal{N}(x_i|\mu_1, \Sigma_1) + \sum_{i=1}^{n} \mathbf{1}\{y = 0\} \ln \mathcal{N}(x_i|\mu_0, \Sigma_0)$$

Learning time !!!!! **MLE**

根据 $\hat{\Theta}_{MLE} = \arg\max_{\Theta} \ln P(D; \Theta)$，假设 Σ_0, Σ_1 已知，设 $M_1 = \{(x_i, y_i)|y_i = 1\}, M_0 = \{(x_i, y_i)|y_i = 0\}$，求 p, μ_0, μ_1。

$$\begin{aligned} \mathcal{L}_D(\Theta) &= \ln P(D; \Theta) \\ \ln \mathcal{N}(x|\mu, \Sigma) &= -\frac{d}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma| - \frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu) \\ \frac{\partial \mathcal{L}_D(\Theta)}{\partial p} &= \sum_{i=1} w^n \left(\frac{y_i}{p} - \frac{1 - y_i}{1 - p}\right) = 0 \\ \frac{\partial \mathcal{L}_D(\Theta)}{\partial \mu_0} &= \sum_{(x_i, y_i) \in M_0} \Sigma^{-1}(x_i - \mu_0) = 0 \\ \frac{\partial \mathcal{L}_D(\Theta)}{\partial \mu_1} &= \sum_{(x_i, y_i) \in M_1} \Sigma^{-1}(x_i - \mu_1) = 0 \end{aligned}$$

解得

$$p = \frac{\sum_{i=1}^{n} y_i}{n}$$
$$\mu_0 = \frac{\sum_{(x_i,y_i)\in M_0} x_i}{|M_0|}$$
$$\mu_1 = \frac{\sum_{(x_i,y_i)\in M_1} x_i}{|M_1|}$$

Infering time !!!

Given $x, y =?$， $p, \mu_0, \mu_1, \Sigma_0, \Sigma_1$ are known.

$$\begin{aligned}
P(y=1|x) &\propto P(y=1)P(x|y=1) \\
&= p\mathcal{N}(x|\mu_1, \Sigma_1) \\
P(y=0|x) &\propto P(y=0)P(x|y=0) \\
&= (1-p)\mathcal{N}(x|\mu_0, \Sigma_0)
\end{aligned}$$

若 $P(y=1|x) > P(y=0|x)$ ，则 $y=1$。

5.5 作业

$$\begin{aligned}
g(x) &= \ln \frac{P(y=1|x)}{P(y=0|x)} \\
&= \ln \frac{p\mathcal{N}(x|\mu_1, \Sigma_1)}{(1-p)\mathcal{N}(x|\mu_0, \Sigma_0)}
\end{aligned}$$

设 $\Sigma_1 = \Sigma_0$, 则 $g(x)$ 是线性平面, 可写作 $g(x) = w^T x + b$, 即 Gaussian Discriminative Analysis(GDA), 求 w, b

$$\begin{aligned}
g(x) &= \ln \frac{p}{1-p} + (\frac{1}{2}((x-\mu_0)^T\Sigma^{-1}(x-\mu_0) - (x-\mu_1)^T\Sigma^{-1}(x-\mu_1))) \\
&= \ln \frac{p}{1-p} + \frac{1}{2}((x^T - {\mu_0}^T)\Sigma^{-1}(x-\mu_0) - (x^T - {\mu_1}^T)\Sigma^{-1}(x-\mu_1)))) \\
&= \ln \frac{p}{1-p} + \frac{1}{2}(x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_0 - {\mu_0}^T\Sigma^{-1}x + {\mu_0}^T\Sigma^{-1}\mu_0 + x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_1 + {\mu_1}^T\Sigma^{-1}x - {\mu_1}^T\Sigma^{-1}\mu_1) \\
&= \ln \frac{p}{1-p} + \frac{1}{2}(2({\mu_1}^T\Sigma^{-1}x - {\mu_0}^T\Sigma^{-1}x) + {\mu_0}^T\Sigma^{-1}\mu_0 - {\mu_1}^T\Sigma^{-1}\mu_1) \\
&= (\mu_1 - \mu_0)^T\Sigma^{-1}x + \frac{1}{2}{\mu_0}^T\Sigma^{-1}\mu_0 - \frac{1}{2}{\mu_1}^T\Sigma^{-1}\mu_1 + \ln p - \ln(1-p)
\end{aligned}$$

即 (结合：w 记得算一下转置，Σ 是对称矩阵等内容)

$$\begin{aligned}
w &= \Sigma^{-1}(\mu_1 - \mu_0) \\
b &= \frac{1}{2}{\mu_0}^T\Sigma^{-1}\mu_0 - \frac{1}{2}{\mu_1}^T\Sigma^{-1}\mu_1 + \ln p - \ln(1-p)
\end{aligned}$$

x 是离散的