1.1 Specific Gaussian naive Bayes classifiers and logistic regression

Consider **a specific class** of Gaussian naive Bayes classifiers where:

- y is a boolean variable following a Bernoulli distribution, with parameter $\pi = P(y=1)$ and thus $P(Y = 0) = 1 - \pi$.
- $\mathbf{x} = [x_1, \ldots, x_D]^T$, with each feature x_i a continuous random variable. For each x_i, $P(x_i|y=k)$ is a Gaussian distribution $\mathcal{N}(\mu_{ik}, \sigma_i)$. Note that σ_i is the standard deviation of the Gaussian distribution, which does not depend on k.
- For all $i \neq j$, x_i and x_j are conditionally independent given y (so called "naive" classifier).

Question: please show that the relationship between a discriminative classifier (say logistic regression) and the above specific class of Gaussian naive Bayes classifiers is precisely the form used by logistic regression.

1. 对于GDA，我们有： $y \sim Bern(\pi)$

$$x_i | y=k \sim \mathcal{N}(\mu_{ik}, \sigma_i),\ k \in \{0,1\}$$

于是有

$$P(y) = \pi^y (1-\pi)^{1-y}$$

$$P(x_i | y=k, \mu_{ik}, \sigma_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot e^{-\frac{(x_i-\mu_{ik})^2}{2\sigma_i^2}} \quad , k \in \{0,1\}$$

由于x_i之间条件独立，有：

$$P(x | y=1) = \prod_{i=1}^{D} P(x_i | y=1)$$

$$P(x | y=0) = \prod_{i=1}^{D} P(x_i | y=0)$$

由贝叶斯公式，有：

$$P(y=1 | x) = \frac{P(x|y=1) \cdot P(y=1)}{P(x|y=0) \cdot P(y=0) + P(x|y=1) \cdot P(y=1)}$$

$$= \frac{1}{1 + \frac{P(x|y=0)P(y=0)}{P(x|y=1)P(y=1)}}$$

$$= \frac{1}{1 + \exp\left\{\ln\left(\frac{P(y=0)}{P(y=1)}\right) + \ln\left(\frac{P(x|y=0)}{P(x|y=1)}\right)\right\}}$$

$$= \frac{1}{1 + \exp\left\{\ln\frac{1-\pi}{\pi} + \sum_{i=1}^{D} \ln\left(\exp\left\{\frac{(x_i-\mu_{i1})^2 - (x_i-\mu_{i0})^2}{2\sigma_i^2}\right\}\right)\right\}}$$

$$= \frac{1}{1 + \exp\left\{\ln\frac{1-\pi}{\pi} + \sum_{i=1}^{D} \frac{(x_i-\mu_{i1})^2 - (x_i-\mu_{i0})^2}{2\sigma_i^2}\right\}}$$

$$= \frac{1}{1 + \exp\left\{\ln\frac{1-\pi}{\pi} + \sum_{i=1}^{D}\left(\frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2} + \frac{(\mu_{i0}-\mu_{i1})}{\sigma_i^2} \cdot x_i\right)\right\}}$$

$$= \frac{1}{1+\exp\{w_0 + \sum_{i=1}^{D} w_i x_i\}}$$

其中：$w_0 = \ln\frac{1-\pi}{\pi} + \sum_{i=1}^{D}\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}$

$w_i = \frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}$

而对于Logistic回归，有：

$$P(y=1|x) = \frac{1}{1+\exp\{w_0 + \sum_{i=1}^{D} w_i x_i\}} = \frac{1}{1+\exp\{w^T x\}}$$

他们的形式完全一致。

1.2 General Gaussian naive Bayes classifiers and logistic regression

Removing the assumption that the standard deviation σ_i of $P(x_i|y=k)$ does not depend on k. That is , for each x_i, $P(x_i|y=k)$ is a Gaussian distribution $\mathcal{N}(\mu_{ik}, \sigma_{ik})$, where $i = 1, \ldots, D$ and $k = 0, 1$.

Question: is the new form of $P(y|\mathbf{x})$ implied by this more general Gaussian naive Bayes classifier still the form used by logistic regression? Derive the new form of $P(y|\mathbf{x})$ to prove your answer.

2. 由1题的步骤，可以得到：

$$P(y=1|x) = \frac{1}{1+\exp\left\{\ln\frac{1-\pi}{\pi} + \sum_{i=1}^{D}\left\{\ln\left|\frac{\sigma_{i1}}{\sigma_{i0}}\right| + \frac{\sigma_{i0}^2-\sigma_{i1}^2}{4\sigma_{i1}^2\sigma_{i0}^2}x_i^2 + \frac{\sigma_{i1}^2\mu_{i0}-\sigma_{i0}^2\mu_{i1}}{2\sigma_{i1}^2\sigma_{i0}^2}x_i + \frac{\sigma_{i1}^2\mu_{i0}^2-\sigma_{i0}^2\mu_{i1}^2}{4\sigma_{i1}^2\sigma_{i0}^2}\right\}\right\}}$$

$$= \frac{1}{1+\exp\{w_0 + \sum_{i=1}^{D}(w_i x_i^2 + w_i' x_i)\}}$$

其中：$w_0 = \ln\frac{1-\pi}{\pi} + \sum_{i=1}^{D}\left(\ln\left|\frac{\sigma_{i1}}{\sigma_{i0}}\right| + \frac{\sigma_{i1}^2\mu_{i0}^2-\sigma_{i0}^2\mu_{i1}^2}{4\sigma_{i1}^2\sigma_{i0}^2}\right)$

$w_i = \sum_{i=1}^{D}\frac{\sigma_{i1}^2\mu_{i0}-\sigma_{i0}^2\mu_{i1}}{2\sigma_{i1}^2\sigma_{i0}^2}$

$w_i' = \frac{\sigma_{i0}^2-\sigma_{i1}^2}{4\sigma_{i1}^2\sigma_{i0}^2}$

于是可以看出，此时的GDA与Logistic回归的表示形式并不相同.

1.3 Gaussian Bayes classifiers and logistic regression

Now, consider the following assumptions for our Gaussian Bayes classifiers (without "naive"):

- y is a boolean variable following a Bernoulli distribution, with parameter $\pi = P(y = 1)$ and thus $P(Y = 0) = 1 - \pi$.
- $\mathbf{x} = [x_1, x_2]^T$, i.e., we only consider two features for each sample, with each feature a continuous random variable. x_1 and x_2 are **not** conditional independent given y. We assume $P(x_1, x_2|y = k)$ is a bivariate Gaussian distribution $\mathcal{N}(\mu_{1k}, \mu_{2k}, \sigma_1, \sigma_2, \rho)$, where μ_{1k} and μ_{2k} are means of x_1 and x_2, σ_1 and σ_2 are standard deviations of x_1 and x_2, and ρ is the correlation between x_1 and x_2. The density of the bivariate Gaussian distribution is:

$$P(x_1, x_2|y = k) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp[-\frac{\sigma_2^2(x_1-\mu_{1k})^2 + \sigma_1^2(x_2-\mu_{2k})^2 - 2\rho\sigma_1\sigma_2(x_1-\mu_{1k})(x_2-\mu_{2k})}{2(1-\rho^2)\sigma_1^2\sigma_2^2}].$$

Question: is the form of $P(y|\mathbf{x})$ implied by such not-so-naive Gaussian Bayes classifiers still the form used by logistic regression? Derive the form of $P(y|\mathbf{x})$ to prove your answer.

3. 关键点 $\ln\left(\frac{P(x|y=0)}{P(x|y=1)}\right) = \ln\left(\frac{P(x_1,x_2|y=0)}{P(x_1,x_2|y=1)}\right)$

$$= \ln\left(\frac{\exp\left\{-\frac{\sigma_2^2(x_1-\mu_{10})^2+\sigma_1^2(x_2-\mu_{20})^2-2\rho\sigma_1\sigma_2(x_1-\mu_{10})(x_2-\mu_{20})}{2(1-\rho^2)\sigma_1^2\sigma_2^2}\right\}}{\exp\left\{-\frac{\sigma_2^2(x_1-\mu_{11})^2+\sigma_1^2(x_2-\mu_{21})^2-2\rho\sigma_1\sigma_2(x_1-\mu_{11})(x_2-\mu_{21})}{2(1-\rho^2)\sigma_1^2\sigma_2^2}\right\}}\right)$$

$$= \frac{1}{2(1-\rho^2)\sigma_1^2\sigma_2^2}\cdot\Big[\sigma_2^2[(x_1-\mu_{11})^2-(x_1-\mu_{10})^2] + \sigma_1^2[(x_2-\mu_{21})^2-(x_2-\mu_{20})^2] - 2\rho\sigma_1\sigma_2[(x_1-\mu_{11})(x_2-\mu_{21})-(x_1-\mu_{10})(x_2-\mu_{20})]\Big]$$

$$= \frac{1}{2(1-\rho^2)\sigma_1^2\sigma_2^2}\Big[(2\sigma_2^2(\mu_{10}-\mu_{11})+2\rho\sigma_1\sigma_2(\mu_{21}-\mu_{20}))x_1 + (2\sigma_1^2(\mu_{20}-\mu_{21})+2\rho\sigma_1\sigma_2(\mu_{11}-\mu_{10}))x_2 + \sigma_2^2(\mu_{11}^2-\mu_{10}^2)+\sigma_1^2(\mu_{21}^2-\mu_{20}^2)+2\rho\sigma_1\sigma_2(\mu_{10}\mu_{20}-\mu_{11}\mu_{21})$$

于是 $P(y=1|x) = \dfrac{1}{1+\exp\{w_0+w_1x_1+w_2x_2\}}$

其中 $w_1 = \dfrac{1}{2(1-\rho^2)\sigma_1^2\sigma_2^2}(2\sigma_2^2(\mu_{10}-\mu_{11})+2\rho\sigma_1\sigma_2(\mu_{21}-\mu_{20}))$

$w_2 = \dfrac{1}{2(1-\rho^2)\sigma_1^2\sigma_2^2}(2\sigma_1^2(\mu_{20}-\mu_{21})+2\rho\sigma_1\sigma_2(\mu_{11}-\mu_{10}))$

$w_0 = \ln\dfrac{1-\pi}{\pi} + \dfrac{1}{2(1-\rho^2)\sigma_1^2\sigma_2^2}[\sigma_2^2(\mu_{11}^2-\mu_{10}^2)+\sigma_1^2(\mu_{21}^2-\mu_{20}^2)+2\rho\sigma_1\sigma_2(\mu_{10}\mu_{20}-\mu_{11}\mu_{21})]$

因此，此时的非朴素贝叶斯GDA的形式仍与Logistic回归一致