GDA

Wednesday, August 12, 2020 9:36 AM

(A) Derivation of GDA model

Comparison with Discriminative algorithm:

In Disc we have to classify $y=0$ or $y=1$ (binary example) given features x. Logistic Reg tried to find a straight line (Decision boundary) that separates the two levels. While, GDA looks at how $y=1$ and $y=0$ are made up of and based the prediction of a example by comparing if it fits $y=1$ or $y=0$ properties.

Disc Algo: Find $p(y|x)$ directly

GDA: Find $p(x|y)$ and $p(y)$ then use them to to find $p(y|x)$

makes strong assumption

e.g., $p(x|y=0)$, $p(x|y=1)$

$p(y)$ → class prior

Use bayes rule, $p(y|x) = \dfrac{p(x|y)\,p(y)}{p(x)}$ → total prob

$$\arg\max_y p(y|x) = \arg\max p(x|y)\,p(y)$$

Model:

$y \sim \text{Binomial}(\phi)$

$y \sim \text{Binomial}(\phi)$

$x|y=0 \sim \mathcal{N}(\mu_0, \Sigma)$

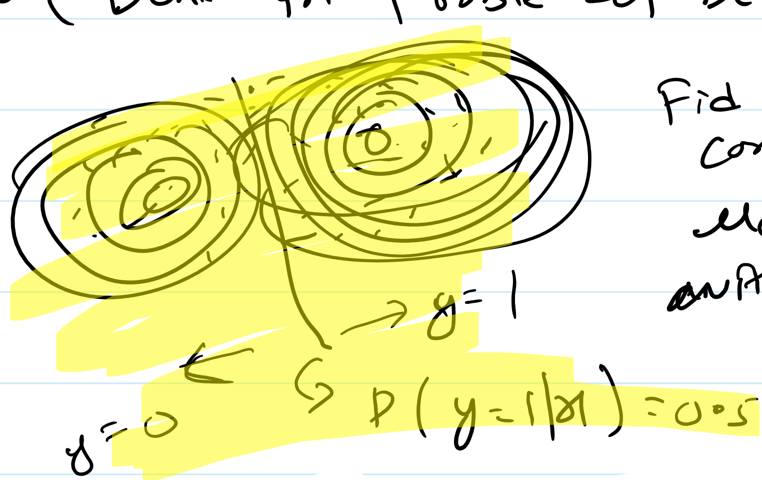
$x|y=1 \sim \mathcal{N}(\mu_1, \Sigma)$

$$\left.\begin{array}{l} p(y) \\ p(x|y=0) \\ p(x|y=1) \end{array}\right\} \rightarrow \text{find prob mass/ density funct}$$

$$\ell(\phi, \mu_0, \mu_1, \Sigma) = \log \prod_{i=1}^{n} p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma)$$

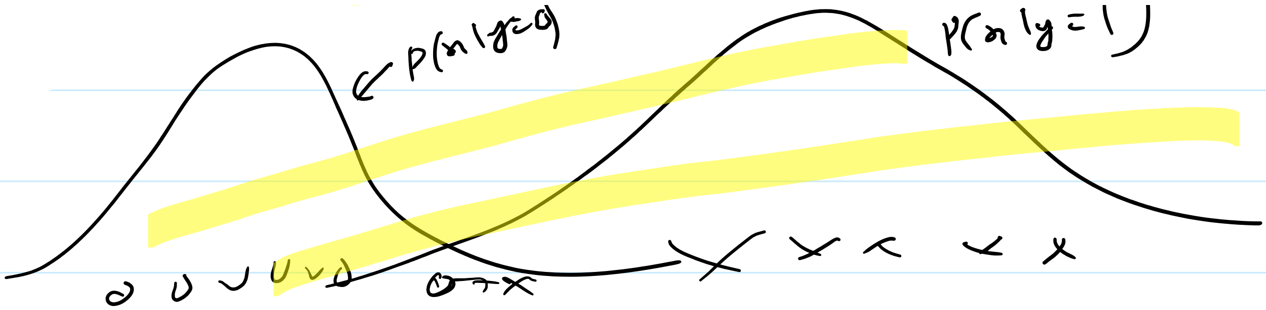
$$= \log \prod_{i=1}^{n} p(x^{(i)} | y^{(i)}; \mu_0, \mu_1, \Sigma)\, p(y^{(i)}; \phi)$$

Take derivative of ℓ w.r.t parameter $(\phi, \mu_0, \mu_1, \Sigma)$ to find the parameter that maximize the likelihood. (Done on Problem Set below)

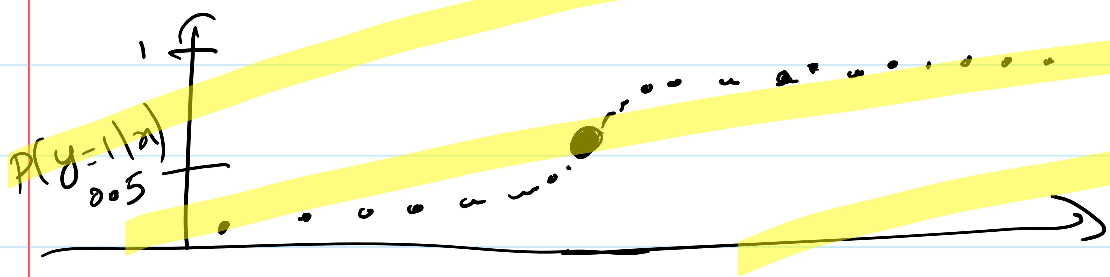


Fit Gaussian controlled by μ_0, μ_1 to shift with shape Σ.

$\boxed{\text{GDA and logistic}}$



$P(y=1) = 0.5$ since data is half and half



Generative

$y \sim \text{Binomial}(\phi)$

$x|y=0 \sim N(\mu_0, \Sigma)$

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

$\Rightarrow$

$\not\Leftarrow$

discriminative (logistic)

$$p(y=1|x) = \frac{1}{1+e^{-\theta^T x}}$$

(B) Problem Set 1 (c) (d) (e)

(c) [5 points] Recall that in GDA we model the joint distribution of (x, y) by the following equations:

$$
\begin{aligned}
p(y) &= \begin{cases} \phi & \text{if } y = 1 \\ 1-\phi & \text{if } y = 0 \end{cases} \\
p(x|y=0) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right) \\
p(x|y=1) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right),
\end{aligned}
$$

where ϕ, μ_0, μ_1, and Σ are the parameters of our model.

Suppose we have already fit ϕ, μ_0, μ_1, and Σ, and now want to predict y given a new point x. To show that GDA results in a classifier that has a linear decision boundary, show the posterior distribution can be written as

$$
p(y = 1 \mid x; \phi, \mu_0, \mu_1, \Sigma) = \frac{1}{1 + \exp(-(\theta^T x + \theta_0))},
$$

where $\theta \in \mathbb{R}^n$ and $\theta_0 \in \mathbb{R}$ are appropriate functions of ϕ, Σ, μ_0, and μ_1.

Answer:

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

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

$$= \frac{1}{1 + \boxed{\frac{\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))\,(1-\phi)}{\exp(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0))\,(\phi)}} \rightarrow x}$$

$$x = e^{\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right) - \left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)}\, e^{\log\left(\frac{1-\phi}{\phi}\right)}$$

$$x = e^{\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)}\, e^{\log\left(\frac{1-\phi}{\phi}\right)}$$

$$= \frac{1}{1+\exp(-[(\mu_1-\mu_0)^T\Sigma^{-1}x) + \frac{1}{2}(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1) - \log\frac{1-\phi}{\phi}]}$$

$$\underbrace{(\mu_1-\mu_0)^T\Sigma^{-1}}_{\theta^T}\ \underbrace{x}_{x}\ \underbrace{\tfrac{1}{2}(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1) - \log\tfrac{1-\phi}{\phi}}_{\theta_0}$$

So we can take

$$\theta = (\mu_1-\mu_0)^T\Sigma^{-1}, \theta_0 = \frac{1}{2}(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1) - \log\frac{1-\phi}{\phi}$$

$\Rightarrow$ for coding part - - - - - - $*$

Using $(A+B)C = (AC) + (BC)$ Also $A^T + B^T = (A+B)^T$

(d) [7 points] For this part of the problem only, you may assume n (the dimension of x) is 1, so that $\Sigma = [\sigma^2]$ is just a real number, and likewise the determinant of Σ is given by $|\Sigma| = \sigma^2$. Given the dataset, we claim that the maximum likelihood estimates of the parameters are given by

$$\begin{aligned}
\phi &= \frac{1}{m}\sum_{i=1}^m 1\{y^{(i)} = 1\} \\
\mu_0 &= \frac{\sum_{i=1}^m 1\{y^{(i)} = 0\}x^{(i)}}{\sum_{i=1}^m 1\{y^{(i)} = 0\}} \\
\mu_1 &= \frac{\sum_{i=1}^m 1\{y^{(i)} = 1\}x^{(i)}}{\sum_{i=1}^m 1\{y^{(i)} = 1\}} \\
\Sigma &= \frac{1}{m}\sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T
\end{aligned}$$

The log-likelihood of the data is

$$\begin{aligned}
\ell(\phi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^m p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma) \\
&= \log \prod_{i=1}^m p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma)p(y^{(i)}; \phi).
\end{aligned}$$

By maximizing ℓ with respect to the four parameters, prove that the maximum likelihood estimates of ϕ, μ_0, μ_1, and Σ are indeed as given in the formulas above. (You may assume that there is at least one positive and one negative example, so that the denominators in the definitions of μ_0 and μ_1 above are non-zero.)

Answer:

$$\ell = \log \prod_{i=1}^m 1\{y^{(i)}=0\}\, P(x^{(i)}|y^{(i)}=0)P(y^{(i)}=0) + 1\{y^{(i)}=1\}P(x^{(i)}|y^{(i)}=1)\,P(y^{(i)}=1)$$

$$= \log \prod_{i=1}^m \left[1\{y^{(i)}=0\}\frac{1}{\sqrt{(2\pi)\sigma^2}}\exp\left(\frac{-(x-\mu_0)^2}{2\sigma^2}\right)\ 1-\phi + \cdots\right]$$

$$= \sum_{i=1}^m \log\left(1\{y^{(i)}=0\}\frac{1}{\sqrt{2\pi}\,\sigma^2}\exp\left(-\frac{(x-\mu_0)^2}{2\sigma^2}\right)(1-\phi) + \cdots\right]$$

$$= \sum_{i=1}^m \log\left(1\{y^{(i)}=0\}\frac{1}{\sqrt{2\pi}\,\sigma^2}\exp\left(-\frac{(x-\mu_0)^2}{2\sigma^2}\right) + 1\{y^{(i)}=1\}\cdots\right) + \log\left(\phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}}\right)$$

$$\frac{\partial \ell}{\partial \phi} = \sum_{i=1}^m \frac{\partial}{\partial \phi}\left[y^{(i)}\log(\phi) + (1-y^{(i)})\log(1-\phi)\right] = 0$$

i.e. $\displaystyle\sum_{i=1}^m \frac{y^i}{\phi} + \frac{1-y^{(i)}}{1-\phi} = 0$

$$\sum_{i=1}^m y^{(i)} - \phi y^{(i)} + \phi - \phi y^{(i)} = 0$$

$$\sum_{i=1}^{m} y^{(i)} + \phi \sum_{i=1}^{m} (1-2y^{(i)}) = 0 \Rightarrow \phi = \frac{\sum_{i=1} y^{(i)}}{\sum_{i=1}^{m} (2y^{(i)}-1)}$$

$$\rightarrow \sum_{i=1}^{m} 1 \left[\text{for any } y=1\right]$$

$$\boxed{\phi = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=1\}}{m}}$$

$$l_{\mu_0} = \frac{\partial}{\partial \mu_0} \sum_{i=1}^{m} \log\left(1\{y^{(i)}=0\} \frac{1}{\sqrt{2\pi}\,\sigma^2} \exp\left(-\frac{(x^{(i)}-\mu_0)^2}{2\sigma^2}\right)\right)$$

$$\boxed{\frac{d}{dx} \log e^{(x-2)^2} = \frac{1}{e^{(x-2)^2}} e^{(x-2)^2} \frac{d(x-2)^2}{dx}}$$

or $\log(e^{x}) = x$ [log and e are inverse function]

$$= \sum_{i=1}^{m} 1\{y^{(i)}=0\} \frac{\partial}{\partial \mu_0} \log\left(e^{-\frac{(x^{(i)}-\mu_0)^2}{2\sigma^2}}\right)$$

$$= \sum_{i=1}^{m} 1\{y^{(i)}=0\} \frac{\partial}{\partial \mu_0}\left(-\frac{(x^{(i)}-\mu_0)^2}{2\sigma^2}\right) = -1\{y^{(i)}=0\} \sum_{i=1}^{m} \frac{x^{(i)}-\mu_0}{\sigma^2} = 0$$

So $-\sum_{i=1}^{m} 1\{y^{(i)}=0\} x^{(i)} + \sum_{i=1}^{m} 1\{y^{(i)}=0\} \mu_0 = 0$

$$\therefore \boxed{\mu_0 = \sum_{i=1}^{m} 1\{y^{(i)}=0\} x^{(i)} \Big/ \sum_{i=1}^{m} 1\{y^{(i)}=0\}}$$

Similarly for μ_1

Similarly for $\Sigma = \sigma^2$,

Note $y = \sigma^2$
$\sqrt{y} = \sigma$

Bad approximation of GDA on data set 1 $\Rightarrow$ Since the dataset is not Gaussian
Fix (take $\log(x^{(i)}$'s))

(c) Implementation Notes

Algorithm

1. Compute $\phi, \mu_0, \mu_1, \Sigma$ using above formula
2. compute θ using '*'

③ Find $P(y=1|x)$; if $P(y=1|x) \geq 0.5$ predict that label is '1' else predict that label is 0.