ML derivations

Taotao Tan

2024-10-11

Here is my personal notes for deriving ML models

- Topic 1: Some useful facts about distributions, linear algebra, etc
- Topic 2: Gaussian Discriminate Model

Topic 1: Some useful facts

Probabilities: Gaussian distribution: $f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

Its negative log-likelihood is $l(\mu, \sigma^2 \mid x) = \frac{(x-\mu)^2}{2\sigma^2} + \log(\sigma) + \log(\sqrt{2\pi})$

Here is the derivatives w.r.t μ, σ^2:

$\frac{\partial l}{\partial \mu} = -\frac{x-\mu}{\sigma^2}$

$\frac{\partial l}{\partial \sigma^2} = \frac{1}{2\sigma^2} - \frac{(x-\mu)^2}{2\sigma^4}$

Multi-variate Gaussian distribution: $f(x \mid \mu, \Sigma) = (2\pi)^{-k/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$

Its negative log-likelihood is $l(\mu, \Sigma \mid x) = \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) + \frac{k}{2}\log(2\pi) + \frac{1}{2}\log(\det(\Sigma))$

Here is the derivatives w.r.t μ, Σ (reference here):

$\frac{\partial l}{\partial \mu} = -\Sigma^{-1}(x-\mu)$

$\frac{\partial l}{\partial \Sigma} = \frac{1}{2}\left(-\Sigma^{-1}(x-\mu)(x-\mu)^T\Sigma^{-1} + \Sigma^{-1}\right)$

Gaussian Discriminate Model

Let $X \in \mathbb{R}^{m \times d}$, $y \in \{0, 1\}$. We assume:

$$\begin{aligned} y_i &\sim \text{Bernoulli}(\phi) \\ x_i \mid y_i = 0 &\sim N(\mu^0, \Sigma) \\ x_i \mid y_i = 1 &\sim N(\mu^1, \Sigma) \end{aligned}$$

For one instance, we can write the joint distribution as:

$$\begin{aligned} P(x_i, y_i) &= P(y_i) \cdot P(x_i \mid y_i) \\ &= \phi^{y_i}(1-\phi)^{1-y_i} \cdot [f_N(\mu^1, \Sigma)]^{y_i} \cdot [f_N(\mu^0, \Sigma)]^{1-y_i} \end{aligned}$$

The log-likelihood can be written as:

$$\begin{aligned} l(\phi, \mu^1, \mu^0, \Sigma \mid (x_i, y_i)) &= \log\left(\phi^{y_i}(1-\phi)^{1-y_i} \cdot [f_N(\mu^1, \Sigma)]^{y_i} \cdot [f_N(\mu^0, \Sigma)]^{1-y_i}\right) \\ &= y_i \log(\phi) + (1-y_i)\log(1-\phi) - \\ &\quad y_i\left(\frac{1}{2}(x_i - \mu^1)^T \Sigma^{-1}(x_i - \mu^1) + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log(\det(\Sigma))\right) - \\ &\quad (1-y_i) \cdot \left(\frac{1}{2}(x_i - \mu^0)^T \Sigma^{-1}(x_i - \mu^0) + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log(\det(\Sigma))\right) \end{aligned}$$

Here I will write the derivative for a single instance. The derivative for the entire dataset $D[x, y]$ is simply the average across each sample.

$$\begin{aligned} \frac{\partial l}{\partial \phi} &= \frac{y_i}{\phi} - \frac{1-y_i}{1-\phi} \\ \frac{\partial l}{\partial \mu^1} &= y_i \Sigma^{-1}(x_i - \mu^1) \\ \frac{\partial l}{\partial \mu^0} &= (1-y_i)\Sigma^{-1}(x_i - \mu^0) \\ \frac{\partial l}{\partial \Sigma} &= -\frac{y_i}{2}[-\Sigma^{-1}(x_i - \mu^1)(x_i - \mu^1)^T \Sigma^{-1} + \Sigma^{-1}] - \frac{1-y_i}{2}[-\Sigma^{-1}(x_i - \mu^0)(x_i - \mu^0)^T \Sigma^{-1} + \Sigma^{-1}] \end{aligned}$$

I will first solve $\frac{1}{m}\sum_{i=1}^m \frac{\partial l}{\partial \phi} = 0$:

$$\begin{aligned} \frac{1}{m}\sum_{i=1}^m \frac{\partial l}{\partial \phi} &= \frac{1}{m}\sum_{i=1}^m \left(\frac{y_i}{\phi} - \frac{1-y_i}{1-\phi}\right) = 0 \\ \phi &= \frac{\sum_{i=1}^m y_i}{m} \end{aligned}$$

Here we solve $\frac{1}{m}\sum_{i=1}^{m}\frac{\partial l}{\partial \mu^1}=0$:

$$\begin{aligned}
\frac{\partial l}{\partial \mu^1} &= \frac{1}{m}\sum_{i=1}^{m}(y_i\Sigma^{-1}(x_i-\mu^1))=0\\
\mu^1 &= (\sum_{i=1}^{m} y_i\Sigma^{-1})^{-1}(\sum_{i=1}^{m} y_i\Sigma^{-1}x_i)\\
&= (\sum_{i=1}^{m} y_i)^{-1}\Sigma^{-1}\Sigma(\sum_{i=1}^{m} y_i x_i)\\
&= \frac{\sum_{i=1}^{m} y_i x_i}{\sum_{i=1}^{m} y_i}\\
&= \frac{\sum_{i=1}^{m}\mathbb{I}(y_i=1)x_i}{\sum_{i=1}^{m}\mathbb{I}(y_i=1)}
\end{aligned}$$

Similarly, we can solve $\frac{1}{m}\sum_{i=1}^{m}\frac{\partial l}{\partial \mu^0}=0$:

$$\begin{aligned}
\frac{\partial l}{\partial \mu^0} &= \frac{1}{m}\sum_{i=1}^{m}((1-y_i)\Sigma^{-1}(x_i-\mu^0))=0\\
\mu^1 &= \frac{\sum_{i=1}^{m}(1-y_i)x_i}{\sum_{i=1}^{m}(1-y_i)}\\
&= \frac{\sum_{i=1}^{m}\mathbb{I}(y_i=0)x_i}{\sum_{i=1}^{m}\mathbb{I}(y_i=0)}
\end{aligned}$$

Then I will need to find the root for $\frac{1}{m}\sum_{i=1}^{m}\frac{\partial l}{\partial \Sigma}=0$. It's a bit complicated, but here are the steps:

$$\frac{1}{m}\sum_{i=1}^{m}\left(-\frac{y_i}{2}[-\Sigma^{-1}(x-\mu^1)(x-\mu^1)^T\Sigma^{-1}+\Sigma^{-1}]\right)=\frac{1}{m}\sum_{i=1}^{m}\left(-\frac{1-y_i}{2}[-\Sigma^{-1}(x-\mu^0)(x-\mu^0)^T\Sigma^{-1}+\Sigma^{-1}]\right)$$

$$\sum_{i=1}^{m}\left(\frac{y_i}{2}\Sigma^{-1}(x-\mu^1)(x-\mu^1)^T\Sigma^{-1}\right)-\sum_{i=1}^{m}\frac{y_i}{2}\Sigma^{-1}=\sum_{i=1}^{m}\left(\frac{y_i-1}{2}\Sigma^{-1}(x-\mu^0)(x-\mu^0)^T\Sigma^{-1}\right)-\sum_{i=1}^{m}\frac{y_i-1}{2}\Sigma^{-1}$$

We might re-arrange the terms, with:

$$\begin{aligned}
&\text{left: } \sum_{i=1}^{m}(\frac{y_i}{2}\Sigma^{-1}(x_i-\mu^1)(x_i-\mu^1)^T\Sigma^{-1})-\sum_{i=1}^{m}(\frac{y_i-1}{2}\Sigma^{-1}(x_i-\mu^0)(x_i-\mu^0)^T\Sigma^{-1})\\
&\text{right: } \sum_{i=1}^{m}(\frac{y_i}{2}\Sigma^{-1})-\sum_{i=1}^{m}(\frac{y_i-1}{2}\Sigma^{-1})
\end{aligned}$$

For the left term, we have the pattern of $\Sigma^{-1}v_1v_1^T\Sigma^{-1}+\Sigma^{-1}v_2v_2^T\Sigma^{-1}+\ldots$. This can be simplified to $\Sigma^{-1}(v_1v_1^T+v_2v_2^T+...)\Sigma^{-1}$. With this rule, we can further reduce the left to:

$$\Sigma^{-1}\left(\sum_{i=1}^{m}\left(\frac{y_i}{2}(x_i-\mu^1)(x_i-\mu^1)^T-\frac{y_i-1}{2}(x_i-\mu^0)(x_i-\mu^0)^T\right)\right)\Sigma^{-1}$$

It's important to notice that $y_i\in\{0,1\}$ for each instance. This allows us to compactly write the expression as:

$$\Sigma^{-1}\left(\sum_{i=1}^{m}\left(\frac{1}{2}(x_i-\mu^{y_i})(x_i-\mu^{y_i})^T\right)\right)\Sigma^{-1}$$

On the right hand side, we can simplify it to:

$$\sum_{i=1}^{m}(\frac{y_i}{2}\Sigma^{-1}) - \sum_{i=1}^{m}(\frac{y_i-1}{2}\Sigma^{-1}) = \frac{m}{2}\Sigma^{-1}$$

Let left = right, then we have:

$$\begin{aligned}
\Sigma^{-1}\left(\sum_{i=1}^{m}\left(\frac{1}{2}(x_i-\mu^{y_i})(x_i-\mu^{y_i})^T\right)\right)\Sigma^{-1} &= \frac{m}{2}\Sigma^{-1} \\
\Sigma\Sigma^{-1}\left(\sum_{i=1}^{m}\left(\frac{1}{2}(x_i-\mu^{y_i})(x_i-\mu^{y_i})^T\right)\right)\Sigma^{-1}\Sigma &= \frac{m}{2}\Sigma\Sigma^{-1}\Sigma \\
\sum_{i=1}^{m}\left(\frac{1}{2}(x_i-\mu^{y_i})(x_i-\mu^{y_i})^T\right) &= \frac{m}{2}\Sigma \\
\Sigma &= \frac{1}{m}\left(\sum_{i=1}^{m}(x_i-\mu^{y_i})(x_i-\mu^{y_i})^T\right)
\end{aligned}$$

In summary, we have the following results for GDA (assume same covariance matrix):

$$\begin{aligned}
\phi &= \frac{\sum_{i=1}^{m} y_i}{m} \\
\mu^1 &= \frac{\sum_{i=1}^{m} y_i x_i}{\sum_{i=1}^{m} y_i} \\
\mu^0 &= \frac{\sum_{i=1}^{m}(1-y_i)x_i}{\sum_{i=1}^{m}(1-y_i)} \\
\Sigma &= \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu^{y_i})(x_i-\mu^{y_i})^T
\end{aligned}$$

It's well known that the GDA is identical to logistic regression. Here is why:

$$\begin{aligned}
P(y=1 \mid x) &= \frac{P(x \mid y=1)\cdot P(y=1)}{P(x \mid y=0)\cdot P(y=0) + P(x \mid y=1)\cdot P(y=1)} \\
&= \frac{1}{\frac{P(x|y=0)\cdot P(y=0)}{P(x|y=1)\cdot P(y=1)} + 1}
\end{aligned}$$

Let's focus on the term $\frac{P(x|y=0)\cdot P(y=0)}{P(x|y=1)\cdot P(y=1)}$ for a moment:

$$\begin{aligned}
\frac{P(x \mid y=0)\cdot P(y=0)}{P(x \mid y=1)\cdot P(y=1)} &= \exp\left(\log\left(\frac{P(x \mid y=0)}{P(x \mid y=1)}\right) + \log\frac{P(y=0)}{y=1}\right) \\
&= \exp\left(\log\left(\frac{\exp[-\frac{1}{2}(x-\mu^0)^T\Sigma^{-1}(x-\mu^0)]}{\exp[-\frac{1}{2}(x-\mu^1)^T\Sigma^{-1}(x-\mu^1)]}\right) + \log\left(\frac{1-\phi}{\phi}\right)\right) \\
&= \exp\left(-\frac{1}{2}[(x-\mu^0)^T\Sigma^{-1}(x-\mu^0) + (x-\mu^1)^T\Sigma^{-1}(x-\mu^1)] + \log\left(\frac{1-\phi}{\phi}\right)\right) \\
&= \exp\left((\mu^0-\mu^1)^T\Sigma^{-1}x + \log(\frac{1-\phi}{\phi}) + \frac{1}{2}(\mu^1-\mu^0)^T\Sigma^{-1}(\mu^1+\mu^0)\right)
\end{aligned}$$

Let $\theta = -(\mu^0-\mu^1)^T\Sigma^{-1}$, $\theta_0 = -\log(\frac{1-\phi}{\phi}) - \frac{1}{2}(\mu^1-\mu^0)^T\Sigma^{-1}(\mu^1+\mu^0)$, then:

$$P(y = 1 \mid x) = \frac{1}{1 + \exp(-\theta x - \theta_0)}$$

which is the same as logistic regression.