S&DS 365 Homework 3 Solutions

Yale University, Department of Statistics

Mar 05, 2021

1 Problem 1:Multi-class Classification and MLE

a)

The loss we want to minimize is

$$R(f) = E[1(f(x) \neq y)]$$

apply conditional probability by conditioning on the features x.

$$E[E[1(f(x) \neq y)|x]] = E[\sum_{j=1}^{k} 1(f(x) \neq j)P(y = j|x)]$$

then, for every value of x we can only assign one class. If we want to minimize the above, we must assign observation x to the class j with the highest conditional probability $P(y = j|x)$, in other words

$$f(x) = \text{argmax}_j P(y = j|x)$$

b)

(Answer does not need to be this thorough, one can just argue the given model defines the same probability distribution.)

Starting from the initial definition of the model

$$\begin{aligned} P(y_i = j|x_i) &= \frac{\exp(x_i^T \beta_j)}{\sum_{m=1}^{k} \exp(x_i^T \beta_m)} \\ &= \prod_{j=1}^{k} \left(\frac{\exp(x_i^T \beta_j)}{\sum_{m=1}^{k} \exp(x_i^T \beta_m)} \right)^{1(y_i=j)} \end{aligned}$$

the exponents are indicators, and only the specific j where $y_i = j$ will have an exponent of 1 and the rest an exponent of 0. Apply exponent rules and noting the denominator is the same in all fractions,

$$\frac{\exp^{\sum_{j=1}^{k} x_i^T \beta_j 1(y_i=j)}}{\sum_{m=1}^{k} \exp(x_i^T \beta_m)} = \frac{\exp(x_i^T \beta_{y_i})}{\sum_{m=1}^{k} \exp(x_i^T \beta_m)}$$

and we take the log likelihood

$$l(B) = \sum_{i=1}^{n} x_i^T \beta_{y_i} - \ln(\sum_{m=1}^{k} \exp(x_i^T \beta_m))$$

c)

We take the argmax of the likelihood

$$\hat{y}_i = \text{argmax}_m P(y_i = j|x_i)$$
$$= \text{argmax}_j \frac{\exp(x_i^T \beta_j)}{\sum_{m=1}^k \exp(x_i^T \beta_m)}$$

the denominator is unaffected by the choice of j so we have

$$\text{argmax}_j \exp(x_i^T \beta_j$$

and the exponential function is an increasing function so the argmax occurs at the largest exponent.

$$\text{argmax}_j x_i^T \beta_j$$

therefore the assigned class is the β_j with largest inner product with the observed point x_i.

d)

$$-\ln P(y_i|x_i) = -\sum_{l=1}^k 1(y_i = l) \ln(p_l(x_i) = -\ln p_{y_i}(x_i)$$

since only the l such that $y_i = l$ will have an indicator $1(y_i = l)$ of 1, the rest are 0 and thus do not appear in the final expression.

2 2: Generative Modeling

a)

We essentially flip a coin y_i, if $y_i = 1$ then x_i has a $N(\mu_1, \Sigma_1)$ distribution and if $y_i = 0$ then x_i has a $N(\mu_0, \Sigma_0)$ distribution. Therefore

$$p(x_i|y_i = 1) = \frac{1}{\sqrt{2\pi \text{det}(\Sigma_1)}} \exp\left(-\frac{(x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1)}{2}\right)$$

b)

The same as above but with μ_0 and Σ_0

$$p(x_i|y_i = 0) = \frac{1}{\sqrt{2\pi \text{det}(\Sigma_0)}} \exp\left(-\frac{(x - \mu_0)^T \Sigma_0^{-1} (x - \mu_0)}{2}\right)$$

c)

c.a)

Here we apply Bayes rule

$$
\begin{aligned}
P(y_i = 1|x_i) &= \frac{P(x_i|y_i = 1)P(y_i = 1)}{P(x_i)} \\
&= \frac{P(x_i|y_i = 1)P(y_i = 1)}{P(x_i|y_i = 1)P(y_i = 1) + P(x_i|y_i = 0)P(y_i = 0)} \\
&= \frac{\pi_1 \left(\frac{1}{\sqrt{2\pi \det(\Sigma_1)}} \exp\left(-\frac{(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)}{2} \right) \right)}{\pi_1 \left(\frac{1}{\sqrt{2\pi \det(\Sigma_1)}} \exp\left(-\frac{(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)}{2} \right) \right) + (1-\pi_1) \left(\frac{1}{\sqrt{2\pi \det(\Sigma_0)}} \exp\left(-\frac{(x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0)}{2} \right) \right)}
\end{aligned}
$$

note this is just the expressions from the previous questions scaled by π_1 and $1 - \pi_1$.

c.b)

Recall the the classifier outputs 1 when $P(y_i = 1|x_i) > P(y_i = 0|x_i)$ and 0 else. The two conditional probabilities $P(y_i = 1|x_i), P(y_i = 0|x_i)$ have the same denominator so we only care which has the larger numerator. We must solve for the values of x_i such that

$$
\begin{aligned}
&\pi_1 \left(\frac{1}{\sqrt{2\pi \det(\Sigma_1)}} \exp\left(-\frac{(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)}{2} \right) \right) \geq (1-\pi_1) \left(\frac{1}{\sqrt{2\pi \det(\Sigma_0)}} \exp\left(-\frac{(x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0)}{2} \right) \right) \\
&\frac{\pi_1}{1-\pi_1} \sqrt{\frac{\det \Sigma_0}{\det \Sigma_1}} \geq \exp\left(-\frac{(x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0)}{2} + \frac{(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)}{2} \right) \\
&2 \ln\left(\frac{\pi_1}{1-\pi_1} \right) + \ln\left(\frac{\det \Sigma_0}{\det \Sigma_1} \right) \geq -(x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0) + (x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) \\
&2 \ln\left(\frac{\pi_1}{1-\pi_1} \right) + \ln\left(\frac{\det \Sigma_0}{\det \Sigma_1} \right) \geq -x^T (\Sigma_0^{-1} - \Sigma_1^{-1}) x^T + x^T (2\Sigma_0^{-1}\mu_0 - 2\Sigma_1^{-1}\mu_1) - \mu_0^T \Sigma_0^{-1} \mu_0 + \mu_1 \Sigma_1^{-1} \mu_1 \\
&x^T (\Sigma_0^{-1} - \Sigma_1^{-1}) x^T - x^T (2\Sigma_0^{-1}\mu_0 - 2\Sigma_1^{-1}\mu_1) \geq -\mu_0^T \Sigma_0^{-1} \mu_0 + \mu_1 \Sigma_1^{-1} \mu_1 - 2 \ln\left(\frac{\pi_1}{1-\pi_1} \right) - \ln\left(\frac{\det \Sigma_0}{\det \Sigma_1} \right)
\end{aligned}
$$

note that this is a quadratic expression in x as desired in the question. We have

$$
\begin{aligned}
A &= \Sigma_0^{-1} - \Sigma_1^{-1} \\
\nu &= -(2\Sigma_0^{-1}\mu_0 - 2\Sigma_1^{-1}\mu_1) \\
\tau &= -\mu_0^T \Sigma_0^{-1} \mu_0 + \mu_1 \Sigma_1^{-1} \mu_1 - 2 \ln\left(\frac{\pi_1}{1-\pi_1} \right) - \ln\left(\frac{\det \Sigma_0}{\det \Sigma_1} \right)
\end{aligned}
$$

c.c)

In dimension $d = 1$ this is a quadratic function, the region where we assign $\hat{y}_i = 1$ is determined by when the parabola is positive or negative.

c.d)

In this case the second degree term A is zero and we have

$$2x^T\Sigma^{-1}(\mu_1-\mu_0) \geq -\mu_0^T\Sigma_0^{-1}\mu_0 + \mu_1\Sigma_1^{-1}\mu_1 - 2\ln\left(\frac{\pi_1}{1-\pi_1}\right)$$

c.e)

In this case it becomes a linear function of x.

3 3: Margin of a Linear Classifier

Part 1:

g is just a scaled version of w and v by definition is an element of the space H where $v^Tw = 0$. Therefore we have (note that $v^Tw = \langle v, w\rangle$ is an inner product)

$$\langle v, g\rangle = \langle v, \langle x_i, w\rangle w\rangle = (\langle x_i, w\rangle)\,\langle v, w\rangle = 0$$

since we can pull constants in front of an inner product.

Part 2)

$$\begin{aligned}\langle e, w\rangle &= \langle x_i - g, w\rangle \\ &= \langle x_i, w\rangle - \langle g, w\rangle \\ &= \langle x_i, w\rangle - \langle\langle x_i, w\rangle w, w\rangle \\ &= \langle x_i, w\rangle - \langle x_i, w\rangle\langle w, w\rangle \\ &= \langle x_i, w\rangle - \langle x_i, w\rangle \\ &= 0\end{aligned}$$

since we assumes w has unit norm therefore $\langle w, w\rangle = 1$.

Part 3)

$$\begin{aligned}\|v - x_i\|^2 &= \|v - (e - g)\|^2 \\ &= \|(v - e) + q\|^2 \\ &= \|v - e\|^2 + \|g\|^2 + 2\langle v - e, g\rangle \\ &= \|v - e\|^2 + \|g\|^2 + 2 * (0)\end{aligned}$$

as we have shown above the remaining inner product will be 0.

Part 4)

$$\delta_i^2 = \min_{v|v\in H} \|v - e\|^2 + \|g\|^2$$

note that g has no influence on this minimum, therefore we minimize by making $\|v - e\|^2$ as small as possible, which is achieved when $v = e$. Therefore δ_i^2 is the square norm of g

$$\begin{aligned}\delta_i^2 &= \|\langle x_i, w\rangle w\|^2 \\ &= (\langle x_i, w\rangle)^2 \|w\|^2\end{aligned}$$

so far we have been assuming $\|w\|^2 = 1$. But now we will instead work with $\frac{w}{\|w\|}$ which is just the normalized version o w.

$$\begin{aligned}&\left(\langle x_i, \frac{w}{\|w\|}\rangle\right)^2 \|\frac{w}{\|w\|}\|^2 \\ &= \left(\frac{\langle x_i, w\rangle}{\|w\|}\right)^2 \\ &= \left(\frac{\langle x_i, w\rangle}{\|w\|}\right)^2 y_i^2\end{aligned}$$

since $y_i \in \{\pm 1\}$.

$$\delta_i^2 = \left(\frac{y_i\langle x_i, w\rangle}{\|w\|}\right)^2$$

and we have

$$|\delta_i| = \left|\frac{\langle_i x, w\rangle}{\|w\|}\right|$$

is the distance from x_i to the hyperplane.