Chapter 4

Statistical learning

References:
- The Elements of Statistical Learning [13] [1]
- Probabilistic Machine Learning: An Introduction [25] [2]
- Pattern Recognition and Machine Learning [3] [3]
- Mathematics of Machine Learning by Prof. Philippe Rigollet (lecture note) [4]
- Statistical Methods for Machine Learning by Larry Wasserman (lecture note) [5]
- An Introduction to Statistical Learning: with Applications in R [15][6]

4.1 Linear Methods for Classification

As explained in [15, Section 4.2] (Why Not Linear Regression?), there are at least two reasons not to perform classification using a regression method:
- a regression method cannot accommodate a qualitative response with more than two classes;
- a regression method will not provide meaningful estimates of $\mathbb{P}(Y|X)$, even with just two classes.

4.1.1 LDA and QDA

Theorem 4.1.1. The true error rate of a classifier h is given by

$$L(h) := \mathbb{P}(h(X) \neq Y).$$

Consider the special case where $Y \in \mathcal{Y} = \{0, 1\}$. Let $r(x) = \mathbb{P}(Y = 1|X = x)$. In this case the Bayes classification rule h^* is given by

$$h^*(x) = \begin{cases} 1, & r(x) > \frac{1}{2} \\ 0, & r(x) \leq \frac{1}{2}. \end{cases}$$

Prove that the Bayes classification rule is optimal, that is, if h is any other classification rule then $L(h^*) \leq L(h)$.

Proof.

[1] https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf
[2] https://probml.github.io/pml-book/book1.html
[3] https://cds.cern.ch/record/998831/files/9780387310732_TOC.pdf
[4] https://ocw.mit.edu/courses/mathematics/18-657-mathematics-of-machine-learning-fall-2015/lecture-notes/MIT18_657F15_LecNote.pdf
[5] https://www.stat.cmu.edu/~larry/=sml/
[6] https://www.statlearning.com/

For a classifier h, we rewrite the true error rate $L(h)$ by

$$\begin{aligned} L(h) &= \mathbb{P}(h(X) \neq Y) = \mathbb{P}(h(X)=1, Y=0) + \mathbb{P}(h(X)=0, Y=1) \\ &= \mathbb{E}(\mathbb{E}(\mathbf{1}_{\{h(X)=1,Y=0\}}|X)) + \mathbb{E}(\mathbb{E}(\mathbf{1}_{\{h(X)=0,Y=1\}}|X)) \end{aligned}$$

where $\mathbf{1}_A$ is the indicator function over a set A and we write $\mathbb{P}(h(X)=1, Y=0) = \mathbb{E}\mathbf{1}_{\{h(X)=1,Y=0\}}$, the second equality is from the disjoint of two events, and the third equality is from the law of total expectation conditioning on X.

Since $h(X)$ is measurable w.r.t X, then we take it away from the inner expectation. So the above equation becomes

$$\begin{aligned} L(h) &= \mathbb{E}(\mathbf{1}_{\{h(X)=0\}}\mathbb{E}(\mathbf{1}_{\{Y=1\}}|X)) + \mathbb{E}(\mathbf{1}_{\{h(X)=1\}}\mathbb{E}(\mathbf{1}_{\{Y=0\}}|X)) \\ &= \mathbb{E}(\mathbf{1}_{\{h(X)=0\}}r(X)) + \mathbb{E}(\mathbf{1}_{\{h(X)=1\}}(1-r(X))) \\ &= \mathbb{E}(\mathbf{1}_{\{h(X)=0\}}r(X) + \mathbf{1}_{\{h(X)=1\}}(1-r(X))) \end{aligned} \tag{4.1}$$

where we rewrite $\mathbb{E}(\mathbf{1}_{\{Y=1\}}|X) = \mathbb{P}(Y=1|X)$ and replace it by $r(X)$ in the second equality.

For a classifier h and the Bayes classifier h^*, using the equality (4.1) we obtain

$$\begin{aligned} L(h) - L(h^*) &= \mathbb{E}(\mathbf{1}_{\{h(X)=0\}}r(X) + \mathbf{1}_{\{h(X)=1\}}(1-r(X))) - \mathbb{E}(\mathbf{1}_{\{h^*(X)=0\}}r(X) + \mathbf{1}_{\{h^*(X)=1\}}(1-r(X))) \\ &= \mathbb{E}[(\mathbf{1}_{\{h(X)=0\}} - \mathbf{1}_{\{h^*(X)=0\}})r(X) + (\mathbf{1}_{\{h(X)=1\}} - \mathbf{1}_{\{h^*(X)=1\}})(1-r(X))] \\ &= \mathbb{E}[(\mathbf{1}_{\{h(X)=0\}} - \mathbf{1}_{\{h^*(X)=0\}})(2r(X)-1)] \end{aligned}$$

where we use identity $\mathbf{1}_{\{h(X)=1\}} = 1 - \mathbf{1}_{\{h(X)=0\}}$ in the third equality.

There are three cases for the last equality. For $h(X) = h^*(X)$, $L(h) - L(h^*) = 0$. For $h(X) = 1, h^*(X) = 0$, $L(h) - L(h^*) = -\mathbb{E}(2r(X)-1) = \mathbb{E}(|2r(X)-1|)$ since $r(X) \leq \frac{1}{2}$. For $h(X) = 0, h^*(X) = 1$, $L(h) - L(h^*) = \mathbb{E}(2r(X)-1)$. Hence, from the above discussion and definition of h^* we have

$$L(h) - L(h^*) = \mathbb{E}[\mathbf{1}_{\{h(X)\neq h^*(X)\}}|2r(X)-1|] \geqslant 0$$

which implies $L(h^*) \leq L(h)$. This gives the desired result. □

Theorem 4.1.2. Suppose that $Y \in \{1, \ldots, k\}$ and $\mathbb{P}(X = x|Y = k)$ is Gaussian $N(\mu_k, \Sigma_k)$.

-
- If $\Sigma_k \neq \Sigma_l$ for any k, l, then the Bayes classifier is

  $$h^*(x) = \operatorname*{argmax}_k \delta_k(x)^{(1)}$$

  provided by

  $$\delta_k^{(1)}(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log(\pi_k).$$

- If $\Sigma_k = \Sigma_l$ for any k, l, then the Bayes classifier is

  $$h^*(x) = \operatorname*{argmax}_k \delta_k^{(2)}(x)$$

  provided by

  $$\delta_k^{(2)}(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k + \log(\pi_k).$$

Exercise 1. If $X|Y=0 \sim N(\mu_0, \Sigma_0)$ and $X|Y=1 \sim N(\mu_1, \Sigma_1)$, then the Bayes rule is

$$h(x) = \begin{cases} 1 & \text{if } r_1^2 < r_2^2 + 2\log\frac{\pi_1}{\pi_0} + \log\frac{|\Sigma_0|}{|\Sigma_1|} \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

where $r_i^2 = (x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i), i = 0, 1$.

Proof. Let $\mathbb{P}(X = x|Y = k) = f_k(x)$ and $\mathbb{P}(Y = k) = \pi_k$ for $k = 0, 1$. From the Bayes' theorem, we have

$$\mathbb{P}(Y = i|X = x) = \frac{f_i(x)\pi_i(x)}{\sum_{k=0}^1 f_k(x)\pi_k}, \quad \text{for } i = 0, 1.$$

Since the Bayes rule is $h^*(x) = \mathbf{1}_{\{\mathbb{P}(Y=1|X=x)>\mathbb{P}(Y=0|X=x)\}}$, we need to simplify $\mathbb{P}(Y = 1|X = x) > \mathbb{P}(Y =$

$0|X = x)$ which is

$$\frac{f_1(x)\pi_1(x)}{\sum_k f_k(x)\pi_k} > \frac{f_0(x)\pi_0(x)}{\sum_k f_k(x)\pi_k}.$$

Note that the denominator can be canceled. Then we have

$$f_1(x)\pi_1(x) > f_0(x)\pi_0(x).$$

Since $X|Y = i \sim \mathcal{N}(\mu_i, \Sigma_i)$ for $i = 0, 1$, the above inequality yields (here we cancel the same term $(2\pi)^{-d/2}$ for both side)

$$|\Sigma_1|^{-1/2} \exp\bigl(-r_1^2/2\bigr) > |\Sigma_0|^{-1/2} \exp\bigl(-r_0^2/2\bigr)$$

where let $r_i^2 = (x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i), i = 0, 1$.

We take logarithm for both side to get

$$-\frac{1}{2}\log|\Sigma_1| - \frac{1}{2}r_1^2 + \log \pi_1 > -\frac{1}{2}\log|\Sigma_0| - \frac{1}{2}r_0^2 + \log \pi_0.$$

This is just

$$r_1^2 < r_0^2 + \log\frac{|\Sigma_0|}{|\Sigma_1|} + 2\log\frac{\pi_1}{\pi_0}.$$

Hence,

$$h^*(x) = \begin{cases} 1, & \text{if } r_1^2 < r_0^2 + \log\frac{|\Sigma_0|}{|\Sigma_1|} + 2\log\frac{\pi_1}{\pi_0}, \\ 0, & \text{otherwise} \end{cases}$$

where let $r_i^2 = (x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i), i = 0, 1$. □

Exercise 2. Consider a classifier with class conditional densities of the form $N(x|\mu_c, \Sigma_c)$. In LDA, we assume $\Sigma_c = \Sigma$ and in QDA, each Σ_c is arbitrary. Assume that $\Sigma_1 = k\Sigma_2$ for $k > 1$. That is, the Gaussian ellipsoids have the same"shape", but the one for class 1 is "wider". Derive an expression for the decision boundary.

Proof. Here we consider two classes that $Y \in \{1, 2\}$ and We use same notations as class. Let $f_k(x) := \mathbb{P}(X = x|Y = k)$ for $k = 1, 2$. Since class conditional densities of $f_k(x)$ are of the form $\mathcal{N}(x|\mu_c, \Sigma_c)$, which are given by

$$f_k(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\Bigl(-\frac{1}{2}(x - \mu_k)^T \Sigma^{-1}(x - \mu_k)\Bigr),\ k = 1, 2.$$

In this question, we consider the decision boundary

$$D(h) = \{x : \mathbb{P}(Y = 1|X = x) = \mathbb{P}(Y = 2|X = x)\}.$$

From the Bayes' theorem, we have

$$\mathbb{P}(Y = i|X = x) = \frac{f_i(x)\pi_i(x)}{\sum_{k=1}^2 f_k(x)\pi_k}, \quad \text{for } i = 1, 2.$$

Using the above equation, the conditional probability equation in decision boundary becomes

$$f_1(x)\pi_1 = f_2(x)\pi_2. \tag{4.3}$$

Plug class conditional densities of $f_k(x)$ into (4.3) and take logarithm for both side, we obtain

$$-\frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(x - \mu_1)^T \Sigma_1^{-1}(x - \mu_1) + \log \pi_1 = -\frac{1}{2}\log|\Sigma_2| - \frac{1}{2}(x - \mu_2)^T \Sigma_2^{-1}(x - \mu_2) + \log \pi_2.$$

Since we know that $\Sigma_1 = k\Sigma_2$ for $k > 1$, the above equation becomes

$$\log\frac{|k\Sigma_2|}{|\Sigma_2|} + (x - \mu_1)^T (k\Sigma_2)^{-1}(x - \mu_1) - (x - \mu_2)^T \Sigma_2^{-1}(x - \mu_2) + 2\log\frac{\pi_2}{\pi_1} = 0.$$

Using $|k\Sigma_2| = k^d|\Sigma_2|$ and expanding the above bracket, we get

$$(\frac{1}{k} - 1)x^T \Sigma_2^{-1} x + (2\mu_2^T - \frac{2}{k}\mu_1^T)\Sigma_2^{-1} x + \frac{1}{k}\mu_1^T \Sigma_2^{-1}\mu_1 - \mu_2^T \Sigma_2^{-1}\mu_2 + d\log k + 2\log\frac{\pi_2}{\pi_1} = 0.$$

□

Exercise 3. Ex 4.2 in [13].

Proof.
part (a)

We follow the same notations as class. Since there are two classes, assume that $Y \in \{1,2\}$. In LDA, let $\mathbb{P}(X=x|Y=k)=f_k(x)$ and $\mathbb{P}(Y=k)=\pi_k$ for $k=1,2$. We need to compare $\mathbb{P}(Y=1|X=x)$ and $\mathbb{P}(Y=2|X=x)$ in LDA. From the Bayes' theorem, we have

$$\mathbb{P}(Y=i|X=x)=\frac{f_i(x)\pi_i(x)}{\sum_{k=1}^{2} f_k(x)\pi_k}, \text{ for } i=1,2.$$

To compare $\mathbb{P}(Y=2|X=x)>\mathbb{P}(Y=1|X=x)$ is equivalent to

$$\frac{f_2(x)\pi_2(x)}{\sum_k f_k(x)\pi_k} > \frac{f_1(x)\pi_1(x)}{\sum_k f_k(x)\pi_k}.$$

Note that the denominator can be canceled. Then we have

$$f_2(x)\pi_2(x) > f_1(x)\pi_1(x). \tag{4.4}$$

Since each class density $f_k(x)$ is multivariate Gaussian, then

$$f_k(x)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\Big(-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\Big), \, k=1,2$$

where two classes have same covariance matrix Σ.

We plug densities of f_k into (4.4) and take logarithm for both side to get

$$-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)+\log\pi_1 > -\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)+\log\pi_2.$$

We expand the above and cancel $x^T\Sigma^{-1}x$, then the above inequality yields

$$-\frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2+\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1+(\mu_2-\mu_1)^T\Sigma^{-1}x+\log\pi_2-\log\pi_1>0$$

Note that we estimate $\pi_1=\frac{n_1}{n}$ and $\pi_2=\frac{n_2}{n}$ since the size of class 1 and class 2 are n_1 and n_2 respectively. Using this estimate, we obtain

$$-\frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2+\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1+(\mu_2-\mu_1)^T\Sigma^{-1}x+\log\Big(\frac{n_2}{n}\Big)-\log\Big(\frac{n_1}{n}\Big)>0$$

Hence,

$$x^T\Sigma^{-1}(\mu_2-\mu_1)>\frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2-\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1+\log\Big(\frac{n_1}{n}\Big)-\log\Big(\frac{n_2}{n}\Big). \tag{4.5}$$

part (b) We label class 1 as C_1 of size n_1 and class 1 as C_2 of size n_2. To minimize the least squares $\sum_{i=1}^n(y_i-\beta_0-\beta^Tx_i)^2$, it suffices to take the derivatives with respect to β_0 and β to zero. We obtain

$$\sum_{i=1}^n(y_i-\beta_0-\beta^Tx_i)=0 \tag{4.6}$$

and

$$\sum_{i=1}^n(y_i-\beta_0-\beta^Tx_i)x_i=0 \tag{4.7}$$

So we just need to solve β_0 and β from above equations.

Note that the target coded of y_i as $-n/n_1$ for class 1 and n/n_2 for class 2, we have

$$\sum_{i=1}^n y_i=-n_1\frac{n}{n_1}+n_2\frac{n}{n_2}=0. \tag{4.8}$$

Plug (4.8) into (4.6), we obtain

$$n\beta_0+\beta^T\sum_{i=1}^n x_i=0 \tag{4.9}$$

Note that

$$\frac{1}{n}\sum_{i=1}^n x_i=\frac{1}{n}(n_1\widehat{\mu}_1+n_2\widehat{\mu}_2). \tag{4.10}$$

Using (4.10), the equation (4.9) becomes

$$\beta_0=(-\frac{n_1}{n}\widehat{\mu}_1^T-\frac{n_2}{n}\widehat{\mu}_2^T)\beta. \tag{4.11}$$

Next, we try to solve β from equation (4.7). Before that, we need some preparation. Since there are two classes, we estimate the mean as in [13, Chapter 4.3] given by

$$\widehat{\mu}_1 = \frac{\sum_{i\in C_1} x_i}{n_1},\ \widehat{\mu}_2 = \frac{\sum_{i\in C_2} x_i}{n_2}.$$

where $i \in C_1$ means that y_i is labeled in the first class coded as $-n/n_1$ and $i \in C_2$ means that y_i is labeled in the second class coded as n/n_2.

Then We have

$$\sum_i x_i = \sum_{i\in C_1} x_i + \sum_{i\in C_2} x_i = n_1\widehat{\mu}_1 + n_2\widehat{\mu}_2. \tag{4.12}$$

Also, We estimate the covariance matrix from our training data as in [13, Chapter 4.3]

$$\widehat{\Sigma} = \frac{1}{n-2}\left[\sum_{i\in C_1}(x_i-\widehat{\mu}_1)(x_i-\widehat{\mu}_1)^T + \sum_{i\in C_2}(x_i-\widehat{\mu}_2)(x_i-\widehat{\mu}_2)^T\right] = \frac{1}{n-2}\left[\sum_{i=1}^n x_i x_i^T - n_1\widehat{\mu}_1\widehat{\mu}_1^T - n_2\widehat{\mu}_2\widehat{\mu}_2^T\right]. \tag{4.13}$$

So

$$\sum_{i=1}^n x_i x_i^T = (n-2)\widehat{\Sigma} + n_1\widehat{\mu}_1\widehat{\mu}_1^T + n_2\widehat{\mu}_2\widehat{\mu}_2^T. \tag{4.14}$$

Moreover, we use the target coded of y_i again to get

$$\sum_{i=1}^n x_i y_i = \sum_{i\in C_1} x_i y_i + \sum_{i\in C_2} x_i y_i = -\frac{n}{n_1}\sum_{i\in C_1} x_i + \frac{n}{n_2}\sum_{i\in C_2} x_i = -n\widehat{\mu}_1 + n\widehat{\mu}_2. \tag{4.15}$$

Now we plug (4.11) into equation (4.7) and use equations (4.12), (4.14), and (4.15) for equation (4.7). Thus, we have

$$(n_1\widehat{\mu}_1 + n_2\widehat{\mu}_2)(-\frac{n_1}{n}\widehat{\mu}_1^T - \frac{n_2}{n}\widehat{\mu}_2^T)\beta + ((n-2)\widehat{\Sigma} + n_1\widehat{\mu}_1\widehat{\mu}_1^T + n_2\widehat{\mu}_2\widehat{\mu}_2^T)\beta = n(\widehat{\mu}_2 - \widehat{\mu}_1). \tag{4.16}$$

After some algebra for LHS of (4.16), note that

$$\begin{aligned}(n_1\widehat{\mu}_1 + n_2\widehat{\mu}_2)(-\frac{n_1}{n}\widehat{\mu}_1^T - \frac{n_2}{n}\widehat{\mu}_2^T) + n_1\widehat{\mu}_1\widehat{\mu}_1^T + n_2\widehat{\mu}_2\widehat{\mu}_2^T &= \frac{n_1n_2}{n}\widehat{\mu}_1\widehat{\mu}_1^T + \frac{n_1n_2}{n}\widehat{\mu}_2\widehat{\mu}_2^T - 2\frac{n_1n_2}{n}\widehat{\mu}_1\widehat{\mu}_2^T \\ &= \frac{n_1n_2}{n}(\widehat{\mu}_1\widehat{\mu}_1^T - 2\widehat{\mu}_1\widehat{\mu}_2^T + \widehat{\mu}_2\widehat{\mu}_2^T) \\ &= \frac{n_1n_2}{n}(\widehat{\mu}_1 - \widehat{\mu}_2)(\widehat{\mu}_1 - \widehat{\mu}_2)^T.\end{aligned}$$

Hence, equation (4.16) becomes

$$(\frac{n_1n_2}{n}(\widehat{\mu}_1 - \widehat{\mu}_2)(\widehat{\mu}_1 - \widehat{\mu}_2)^T + (n-2)\widehat{\Sigma})\beta = n(\widehat{\mu}_2 - \widehat{\mu}_1).$$

Hence,

$$(\frac{n_1n_2}{n}\widehat{\Sigma}_B + (n-2)\widehat{\Sigma})\beta = n(\widehat{\mu}_2 - \widehat{\mu}_1) \tag{4.17}$$

where $\widehat{\Sigma}_B = (\widehat{\mu}_2 - \widehat{\mu}_1)(\widehat{\mu}_2 - \widehat{\mu}_1)^T$. This gives the desired result.

part (c) Since $\widehat{\Sigma}_B\beta = (\widehat{\mu}_2 - \widehat{\mu}_1)(\widehat{\mu}_2 - \widehat{\mu}_1)^T\beta$ and $(\widehat{\mu}_2 - \widehat{\mu}_1)^T\beta$ is a scalar, then $\widehat{\Sigma}_B\beta$ is in the direction of $(\widehat{\mu}_2 - \widehat{\mu}_1)$. Note that equation (4.17) can be rewritten as

$$(n-2)\widehat{\Sigma}\beta = n(\widehat{\mu}_2 - \widehat{\mu}_1) - \frac{n_1n_2}{n}\widehat{\Sigma}_B\beta. \tag{4.18}$$

Since terms $\frac{n_1n_2}{n}\widehat{\Sigma}_B\beta$ and $n(\widehat{\mu}_2 - \widehat{\mu}_1)$ are in the direction of $(\widehat{\mu}_2 - \widehat{\mu}_1)$, then the RHS of (4.18) is also in the direction of $(\widehat{\mu}_2 - \widehat{\mu}_1)$. Thus, β is proportional to $\widehat{\Sigma}^{-1}(\widehat{\mu}_2 - \widehat{\mu}_1)$. From equation (4.5), the least squares regression coefficient is identical to the LDA coefficient up to a scalar multiple.

□

Exercise 4. Show that the Naive Bayes Classifier is equivalent to a linear classification rule.

Proof. See https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote05.html.

□

4.1.2 Logistic regression

Exercise 5. Ex 4.4 in [13] for the multi-class logistic regression model.

Proof. For multi-classes logistic regression model, assume that we have K classes and N labels. The response y_{il} is given by that if the data point x_i is from class l where $1 \le l \le K-1$, then the $l-$th element of y_i is one and others are zero, and if x_i is from class K, then all elements of y_i are zero. So response y_{il} form a target matrix corresponding to sample $1 \le n \le N$ and class $1 \le k \le K$. That is

$$y_i = \mathbf{1}_{\{x \text{ is from class } l \text{ and } i=l\}}.$$

From textbook [13, Section 4.4], we know that the posterior probability that x_i comes from class k are given by

$$\mathbb{P}(y=k|X=x) = \frac{\exp\big(\beta_{k0}+\beta_k^T x\big)}{1+\sum_{l=1}^{K-1}\exp\big(\beta_{l0}+\beta_l^T x\big)},\ k=1,2,\ldots,K-1,$$
$$\mathbb{P}(y=K|X=x) = \frac{1}{1+\sum_{l=1}^{K-1}\exp\big(\beta_{l0}+\beta_l^T x\big)}.$$

Let

$$h_k(x) = \mathbb{P}(y=k|X=x),\ k=1,2,\ldots,K$$

The likelihood function for a data point x is given by

$$L(\beta;x) = h_1(x)^{y_1}h_2(x)^{y_2}\cdots h_{K-1}(x)^{y_{K-1}}(h_K(x))^{1-\sum_{l=1}^{K-1}y_l} \tag{4.19}$$

From the posterior probability $\mathbb{P}(y=k|X=x)$, we have the log-likelihold function for a data point x

$$\ell(\beta;x) = y_1(\beta_{10}+\beta_1^T x)+y_2(\beta_{20}+\beta_2^T x)+\cdots+y_{K-1}(\beta_{(K-1)0}+\beta_{K-1}^T x)+\log(h_K) \tag{4.20}$$

Then sum over the equation (4.20) for all data points x_i, we get the log-likelihood of parameter β, that is,

$$\begin{aligned}\ell(\beta) &= \sum_{i=1}^N\sum_{l=1}^{K-1}[y_{il}\beta_l^T x_i + \log(h_K)]\\ &= \sum_{i=1}^N\sum_{l=1}^{K-1}[y_{il}\beta_l^T x_i - \log\left(1+\sum_{l=1}^{K-1}\exp\big(\beta_{l0}+\beta_l^T x_i\big)\right)]\end{aligned}$$

where x_i is the $i-$th sample, β_l is a vector of coefficients for the $l-$th class with size $(p+1)$, $\beta = [\beta_1,\beta_2,\ldots,\beta_{K-1}]^T$ is of size $(K-1)(p+1)$.

Next, we compute the derivative of $\ell(\beta)$.

$$\begin{aligned}\frac{\partial\ell(\beta)}{\partial\beta_k} &= \sum_{i=1}^N\left[y_{ik}x_i^T - \frac{\exp\big(\beta_{k0}+\beta_k^T x_i\big)}{1+\sum_{l=1}^{K-1}\exp\big(\beta_{l0}+\beta_l^T x_i\big)}x_i^T\right]\\ &= \sum_{i=1}^N(y_{ik}-\mathbb{P}(y=k|X=x_i))x_i^T\\ &= (y_{ik}-h_k(x_i))x_i^T.\end{aligned}$$

Let $y_l = [y_{1l},y_{2l},\ldots,y_{Nl}]^T$ and $p_l = [h_l(x_1),h_l(x_2),\ldots,h_l(x_N)]^T$. Then we have

$$\frac{\partial\ell(\beta)}{\partial\beta} = \begin{bmatrix} X^T(y_1-h_1)\\ X^T(y_2-h_2)\\ \vdots\\ X^T(y_{K-1}-h_{k-1})\end{bmatrix}$$

where X is the $N\times(p+1)$ matrix of x_i values.

The Hessian matrix of $\ell(\beta)$ is given by

$$\frac{\partial^2\ell(\beta)}{\partial\beta_k\partial\beta_k'^T} = -\sum_{i=1}^N h_k(x_i)h_{k'}(x_i)x_ix_i^T,\ \text{ for } k\neq k'$$

and for $k = k'$ we have

$$\begin{aligned}
\frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_k^T} &= -\sum_{i=1}^{N} \left[\frac{\exp\big(\beta_{k0} + \beta_k^T x_i\big) x_i (1 + \sum_{l=1}^{K-1} \exp\big(\beta_{l0} + \beta_l^T x_i\big)) - \exp\big(\beta_{k0} + \beta_k^T x_i\big)^2 x_i}{(1 + \sum_{l=1}^{K-1} \exp\big(\beta_{l0} + \beta_l^T x_i\big))^2} x_i^T \right] \\
&= -\sum_{i=1}^{N} [(h_k(x_i) x_i - h_k(x_i)^2 x_i) x_i^T] \\
&= -\sum_{i=1}^{N} [h_k(x_i)(1 - h_k(x_i)) x_i x_i^T].
\end{aligned}$$

Write above second order derivative in form of matrix. Let H_k be $N \times N$ diagonal matrices for $k = 1, 2, \ldots, K-1$ with diagonal elements $h_k(x_i)(1 - h_k(x_i)$, $i = 1, 2, \ldots, N$. Then we rewrite the second derivative of $\ell(\beta)$ as $k = k'$

$$\frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_k^T} = -X^T H_k X.$$

Let T_k be $N \times N$ diagonal matrices for $k = 1, 2, \ldots, K-1$ with diagonal elements $h_k(x_i)$, $i = 1, 2, \ldots, N$. Then we rewrite the second derivative of $\ell(\beta)$ as $k \neq k'$

$$\frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_k'^T} = -X^T T_k T_{k'} X.$$

Hence, the Hessian matrix of $\ell(\beta)$ is given by

$$\begin{aligned}
\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} &= \begin{bmatrix} -X^T H_1 X & -X^T T_1 T_2 X & \cdots & -X^T T_1 T_{K-1} X \\ -X^T T_2 T_1 X & -X^T H_2 X & \cdots & -X^T T_2 T_{K-1} X \\ \vdots & & \ddots & \vdots \\ -X^T T_{K-1} T_1 X & -X^T T_{K-1} T_2 X & \cdots & -X^T H_{K-1} X \end{bmatrix} \\
&= -\widehat{X}^T W \widehat{X}
\end{aligned}$$

where $\widehat{X} = X \cdot \mathrm{Id}_{K-1}$, Id_{K-1} is a $(K-1) \times (K-1)$ identity matrix, $\widehat{X}$ is a $(K-1) \times (K-1)$ matrix with each block matrix of size $(p+1) \times (p+1)$, and

$$W = \begin{bmatrix} H_1 & T_1 T_2 & \cdots & T_1 T_{K-1} \\ T_2 T_1 & H_2 & \cdots & T_2 T_{K-1} \\ \vdots & & \ddots & \vdots \\ T_{K-1} T_1 & T_{K-1} T_2 & \cdots & H_{K-1} \end{bmatrix}$$

Now our Newton-Raphson algorithm for maximizing the log-likelihood is given by

$$\beta^{new} = \beta^{old} + (\widehat{X}^T W \widehat{X})^{-1} \widehat{X}^T \begin{bmatrix} (y_1 - h_1) \\ (y_2 - h_2) \\ \vdots \\ (y_{K-1} - h_{k-1}) \end{bmatrix}$$

Let

$$y - h = \begin{bmatrix} (y_1 - h_1) \\ (y_2 - h_2) \\ \vdots \\ (y_{K-1} - h_{k-1}) \end{bmatrix}$$

Hence, the algorithm can be expressed as

$$\beta^{new} \leftarrow (\widehat{X}^T W \widehat{X})^{-1} \widehat{X}^T W (\widehat{X} \beta^{old} + W^{-1}(y - h))$$

So β^{new} is the solution of a non-diagonal weighted least squares problem with a response $(\widehat{X}\beta^{old} + W^{-1}(y - h))$. We can still use the Netwon algorithm as an iteratively reweighted least squares algorithm. Let $z = (\widehat{X}\beta^{old} + W^{-1}(y - h))$. The iteratively reweighted least squares algorithm is as follows. Set $\beta^0 = 0$, update β^{new} by

$$\beta^{new} \leftarrow \operatorname*{argmin}_{\beta} (z - \widehat{X}\beta) W (z - \widehat{X}\beta)$$

However, the Hessian maybe not negative definite, Newton-Raphson update cannot perform effective.

Here we implement a improved Newton-Raphson algorithm from paper [11]. Given an intial value β^0, let λ_1 be the largest eigenvalue of Hessian matrix of $\ell(\beta)$ at β^0 defined by $H(\ell(\beta^0))$. Let ε be the step size and let $\alpha = \lambda_1 + \varepsilon \|\frac{\partial \ell(\beta^0)}{\partial \beta}\|_2$. Define the controlling of Hessian matrix H by

$$H_\alpha(\ell(\beta^0)) = \begin{cases} H(\ell(\beta^0)) - \alpha \cdot \mathrm{Id}, & \text{if } \alpha > 0, \\ H(\ell(\beta^0)), & \text{otherwise} \end{cases}$$

where $H_\alpha(\ell(\beta^0))$ is always negative definite.

Update β^{new} by

$$\beta^{new} = \beta^{old} - H_\alpha^{-1}(\ell(\beta^{old})) \frac{\partial \ell(\beta^{old})}{\partial \beta}$$

where we have computed the Hessian and gradient of $\ell(\beta)$ in form of matrix as before. □

4.1.3 SVM

Exercise 6. Show that if their convex hulls intersect, the two sets of points cannot be linearly separable.

Proof. See Bishop 3.4 in `https://www.cise.ufl.edu/~anand/fa05/hw1sol_fall05.pdf`. □

Exercise 7 (Exercise in [3])**.** In the maximum-margin hyperplane problem, let's τ denotes the value of the margin. Show that

$$\frac{1}{\tau^2} = 2\sum \alpha - \sum_{k=1}^{n} \sum_{j=1}^{n} \alpha_k \alpha_j y_k y_j x_k^T x_j.$$

Bibliography

[1] Jean Barbier. “High-dimensional inference: a statistical mechanics perspective”. In: *arXiv preprint arXiv:2010.14863* (2020).

[2] Jean Barbier, Nicolas Macris. “The adaptive interpolation method: a simple scheme to prove replica formulas in Bayesian inference”. In: *Probability Theory and Related Fields* 174.3-4 (2019), pp. 1133–1185. URL: https://arxiv.org/abs/1705.02780.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] Hong-Bin Chen, Jean-Christophe Mourrat, Jiaming Xia. “Statistical inference of finite-rank tensors”. In: *arXiv preprint arXiv:2104.05360* (2021).

[5] Lenaic Chizat, Francis Bach. “On the global convergence of gradient descent for over-parameterized models using optimal transport”. In: *arXiv preprint arXiv:1805.09545* (2018). URL: https://arxiv.org/pdf/1805.09545.

[6] Marco Cuturi. “Sinkhorn distances: Lightspeed computation of optimal transport”. In: *Advances in neural information processing systems* 26 (2013), pp. 2292–2300. URL: http://papers.neurips.cc/paper/4927-sinkhorn-distances-lightspeed-computation-of-optimal-transport.pdf.

[7] Amir Dembo, Ofer Zeitouni. *Large deviations techniques and applications*. Vol. 38. Stochastic Modelling and Applied Probability. Springer-Verlag, Berlin, 2010, pp. xvi+396. ISBN: 978-3-642-03310-0. DOI: 10.1007/978-3-642-03311-7. URL: https://doi.org/10.1007/978-3-642-03311-7.

[8] Partha S. Dey, Qiang Wu. “Fluctuation results for Multi-species Sherrington-Kirkpatrick model in the replica symmetric regime”. In: (Preprint, arXiv:2012.13381). eprint: arXiv:2012.13381.

[9] Oliver Y Feng et al. “A unifying tutorial on Approximate Message Passing”. In: *arXiv preprint arXiv:2105.02180* (2021).

[10] Sacha Friedli, Yvan Velenik. *Statistical mechanics of lattice systems: a concrete mathematical introduction*. Cambridge University Press, 2017.

[11] Stephen M Goldfeld, Richard E Quandt, Hale F Trotter. “Maximization by quadratic hill-climbing”. In: *Econometrica: Journal of the Econometric Society* (1966), pp. 541–551.

[12] Ziv Goldfeld, Kristjan Greenewald. “Gaussian-smoothed optimal transport: Metric structure and statistical efficiency”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 3327–3337. URL: http://proceedings.mlr.press/v108/goldfeld20a/goldfeld20a.pdf.

[13] Trevor. Hastie. *The elements of statistical learning data mining, inference, and prediction*. eng. 2nd ed. Springer series in statistics. New York: Springer, 2009. ISBN: 9780387848587.

[14] Jan-Christian Hütter, Philippe Rigollet. “Minimax estimation of smooth optimal transport maps”. In: *The Annals of Statistics* 49.2 (2021), pp. 1166–1194. URL: https://arxiv.org/pdf/1905.05828.

[15] Gareth James et al. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013. URL: https://faculty.marshall.usc.edu/gareth-james/ISL/.

[16] Ivan Kobyzev, Simon Prince, Marcus Brubaker. “Normalizing flows: An introduction and review of current methods”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020). URL: https://arxiv.org/pdf/1908.09257.

[17] Zhifeng Kong, Kamalika Chaudhuri. "The expressive power of a class of normalizing flow models". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 3599–3609. URL: http://proceedings.mlr.press/v108/kong20a/kong20a.pdf.

[18] Flavien Léger. "A gradient descent perspective on Sinkhorn". In: *Applied Mathematics & Optimization* (2020), pp. 1–13. URL: https://link.springer.com/content/pdf/10.1007/s00245-020-09697-w.pdf.

[19] Marc Lelarge, Léo Miolane. "Fundamental limits of symmetric low-rank matrix estimation". In: *Probability Theory and Related Fields* 173.3-4 (2019), pp. 859–929. URL: https://arxiv.org/abs/1611.03888.

[20] Youssef Marzouk et al. "An introduction to sampling via measure transport". In: *arXiv preprint arXiv:1602.05023* (2016). URL: https://arxiv.org/pdf/1602.05023.

[21] Marc Mezard, Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[22] Andrea Montanari. *Mean field asymptotics in high-dimensional statistics: A few references*. 2020.

[23] J-C Mourrat. "Hamilton–Jacobi equations for finite-rank matrix inference". In: *Annals of Applied Probability* 30.5 (2020), pp. 2234–2260.

[24] Jean-Christophe Mourrat. "Hamilton–Jacobi equations for mean-field disordered systems". In: *Annales Henri Lebesgue* 4 (2021), pp. 453–484.

[25] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL: probml.ai.

[26] Victor M Panaretos, Yoav Zemel. "Statistical aspects of Wasserstein distances". In: *Annual review of statistics and its application* 6 (2019), pp. 405–431. URL: https://arxiv.org/pdf/1806.05500.

[27] Dmitry Panchenko. *The Sherrington-Kirkpatrick model*. Springer Monographs in Mathematics. Springer, New York, 2013, pp. xii+156. ISBN: 978-1-4614-6288-0; 978-1-4614-6289-7. DOI: 10.1007/978-1-4614-6289-7. URL: https://doi-org.proxy2.library.illinois.edu/10.1007/978-1-4614-6289-7.

[28] Filippo Santambrogio. "Optimal transport for applied mathematicians". In: *Birkäuser, NY* 55.58-63 (2015), p. 94. URL: http://math.univ-lyon1.fr/~santambrogio/OTAM-cvgmt.pdf.

[29] Alessio Spantini, Daniele Bigoni, Youssef Marzouk. "Inference via low-dimensional couplings". In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 2639–2709. URL: https://www.jmlr.org/papers/volume19/17-747/17-747.pdf.

[30] Michel Talagrand. *Mean field models for spin glasses. Volume I*. Vol. 54. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics. Basic examples. Springer-Verlag, Berlin, 2011, pp. xviii+485. ISBN: 978-3-642-15201-6. DOI: 10.1007/978-3-642-15202-3. URL: https://doi-org.proxy2.library.illinois.edu/10.1007/978-3-642-15202-3.

[31] Michel Talagrand. *Mean field models for spin glasses. Volume II*. Vol. 55. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics. Advanced replica-symmetry and low temperature. Springer, Heidelberg, 2011, pp. xii+629. ISBN: 978-3-642-22252-8; 978-3-642-22253-5.

[32] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. DOI: 10.1017/9781108627771.