Fun facts in mathematics

Soheil Hajian

March 3, 2024

Contents

1 (Numerical) linear algebra

1.1 Rayleigh quotient

Rayleigh quotient is a quotient defined for symmetric matrices and can be used as a technique for estimating minimum and maximum eigenvalues of a symmetric matrix. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix, i.e., $A = A^\top$. Then the Rayleigh quotient is defined as

$$R(A, \mathbf{w}) := \frac{\mathbf{w}^\top A\mathbf{w}}{\mathbf{w}^\top \mathbf{w}}, \quad \forall \mathbf{w} \in \mathbb{R}^n. \tag{1}$$

It can be shown for a symmetric matrix, A, and its eigenvalues $\{\lambda\}_{i=1}^n$, we have

$$\lambda_{\max} = \max_{\mathbf{w}\in\mathbb{R}^n} R(A, \mathbf{w}), \lambda_{\min} = \min_{\mathbf{w}\in\mathbb{R}^n} R(A, \mathbf{w}), \tag{2}$$

and therefore

$$\min_{\mathbf{w}\in\mathbb{R}^n} R(A, \mathbf{w}) \leq \lambda_i \leq \max_{\mathbf{w}\in\mathbb{R}^n} R(A, \mathbf{w}), \quad \forall i = 1, \ldots, n. \tag{3}$$

2 Probability theory and statistics

2.1 Terminologies

In this section we introduce some terminologies used in statistics.

2.1.1 Statistic and estimators

A statistic or sample statistic is a quantity that is computed using a sample of a population. An estimator is a statistic used to estimate a *quantity of interest* of the population. For example the mean of a sample from a population is a statistic that can be used as an estimator for the mean of the population (more precisely, mean of its probability distribution).

In the following we will introduce properties of the estimators. Suppose we are interested in knowing a quantity of interest which we denote by θ of a population which is described through a random variable, X. For example the average height in a population where the height of an individual in the population is a random variable denoted by X. In this example $\theta := \mathbb{E}[X]$.

Let us denote an estimator of θ by $\hat{\theta}(X)$ (note that $\hat{\theta}(X)$ is stochastic but given an observation x, $\hat{\theta}(x)$ is a fixed value). Then the estimator $\hat{\theta}$ is said to be:

- Unbiased if $\mathbb{E}[\hat{\theta}] = \theta$. In our example, the estimator $\hat{\theta}(X) = \frac{1}{N}\sum_{i=1}^{N} X_i$ where $\{X_i\}_{i=1}^{N}$ are i.i.d. random samples whose probability densities is same as X is an unbiased estimator of $\mathbb{E}[X]$.

- Minimum variance unbiased estimator: if the estimator is unbiased and its variance $\mathbb{V}[\hat{\theta}] := \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$ is the smallest among all unbiased estimators.

- Consistent: if a sequence of estimators $\{\hat{\theta}_n\}_{n=1}^{\infty}$ converges to the quantity of the interest when $n \to \infty$. That is for all $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}\big[|\hat{\theta}_n - \theta| < \varepsilon\big] = 1. \tag{4}$$

2.2 Chebyshev inequality

Let X be a continuous random variable with density function $f_X(x)$, mean μ and standard deviation σ. Then the following inequality holds

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \quad \forall k \geq 1. \tag{5}$$

In order to prove (5) we use definition of the probability and standard deviation of X:

$$\begin{aligned} \mathbb{P}(|X - \mu| \geq k\sigma) &= \int_{|x-\mu|\geq k\sigma} f_X(x)\mathrm{d}x \\ &\leq \int_{|x-\mu|\geq k\sigma} \frac{|x-\mu|^2}{(k\sigma)^2} f_X(x)\mathrm{d}x \\ &\leq \int_{\mathbb{R}} \frac{|x-\mu|^2}{(k\sigma)^2} f_X(x)\mathrm{d}x \\ &= \frac{\sigma^2}{(k\sigma)^2} \\ &= \frac{1}{k^2}. \end{aligned} \tag{6}$$

2.3 Law of large numbers

Let us consider a sequence of independent and identically distributed (i.i.d.) samples $(X_1, X_2, \ldots, X_n)$ from a common random variable X. The law of large numbers states that the mean of the above sequence converges to the mean of X as n grows to infinity. That is

$$\bar{X}_n \to \mu \text{ as } n \to \infty, \tag{7}$$

where

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}, \tag{8}$$

and $\mu = \mathbb{E}(X)$. Convergence should be understood in the sense of almost surely (a.s.) which corresponds to the strong law of large numbers. The weak law of large numbers corresponds to the convergence in probability of the mean of the samples to the mean of X.

Here we prove the weak law of large numbers for the case when X is continuous random variable. From Chebyshev inequality (5) we have for the random variable $\bar{X}_n$

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2_{\bar{X}_n}}{\varepsilon^2}, \tag{9}$$

where we set $\varepsilon = k\sigma_{\bar{X}_n}$. On the other hand from the definition of $\bar{X}_i$ we can conclude that

$$\sigma_{\bar{X}_n} = \frac{1}{\sqrt{n}}\sigma_X.$$

If σ_X is finite we can conclude that

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2_X}{n\varepsilon^2},$$

and therefore

$$1 - \frac{\sigma^2_X}{n\varepsilon^2} \leq \mathbb{P}(|\bar{X}_n - \mu| \leq \varepsilon) \leq 1, \quad \forall \varepsilon > 0. \tag{10}$$

Letting $n \to 0$ yields that $\mathbb{P}(|\bar{X}_n - \mu| \leq \varepsilon) \to 1$ for all $\varepsilon > 0$.

3 Pattern recognition

3.1 Linear discriminant analysis

In this section we introduce a method to find a subspace that that aims to maximize the distance between data points belonging to different classes (intra-classes distance) while minimizing the distance of data points belonging to the same class (inter-classes).

Let us denote the dataset by the tuple (X, y) where $X \in \mathbb{R}^{n\times p}$ is the so-called feature matrix, and $y \in \mathbb{B}^p$ is the target vector. Here p denotes the number of data points in the dataset and n is the dimension of the feature space.

Each column of X corresponds to the features of a data point which we denote by $x^{(j)} \in \mathbb{R}^n$, and similarly each entry of y, i.e., $y^{(j)}$ corresponds to the class of the data point j, for all $j = 1, \ldots, p$. We consider that each data point can belong to one and only one class, and we denote the total number of classes by K. That is $y^{(j)} \in \{1, \ldots, K\}$ for all $j = 1, \ldots, p$.

3.1.1 Scatter matrices and spread

Let us denote the projection of the features of a data point, $x^{(j)}$, into a normalized vector $\mathsf{q} \in \mathbb{R}^n$ by $z^{(j)}$, and for all data points by $z := \mathsf{q}^\top X$. A measure of the spread of projected values can be define by the Euclidean 2-norm, $\|z\|_2$:

$$\|z\|_2 = \mathsf{q}^\top X X^\top \mathsf{q}. \tag{11}$$

The matrix $X^\top X$ is called the scatter matrix and we denote it by $S \in \mathbb{R}^{n \times n}$. Note that from the definition of S we can conclude that S is symmetric and at least positive semi-definite. Therefore maximizing the spread corresponds to finding the normalized vector q that maximizes $\mathsf{q}^\top S \mathsf{q}$, which by Rayleigh quotient (see Section 1.1) corresponds to the eigenvector of S corresponding to the maximum eigenvalue of S.

We will now define two metrics: intra-class and inter-class spread of the dataset. It is convenient to define first the set of data points that belong to the same class. Let us define the set of data points belonging to the same class by

$$\mathcal{D}_k := \big\{ j : y^{(j)} = k, \quad \forall j = 1, \ldots, p \big\}, \tag{12}$$

for all $k = 1, \ldots, K$. This then motivates to rearrange the feature matrix, X, such that data points belonging to the same class be adjacent to each other:

$$X = [X_1, X_2, \ldots, X_K], \tag{13}$$

and similarly the target vector $y = (y_1, \ldots, y_K)^\top$. Note that each matrix X_k can have a different size, i.e., $X_k \in \mathbb{R}^{n \times p_k}$ for all $k = 1, \ldots, K$.

In order to define the intra-class spread, we first translate each class the data points to the origin. The center of a class is defined by

$$c_k := \frac{1}{|\mathcal{D}_k|} \sum_{x \in \mathcal{D}_k} x. \tag{14}$$

We then define the centered feature matrix of each class by

$$X_{k,c} := \Big[x^{(j_1)} - c_k, \quad x^{(j_2)} - c_k, \quad \ldots, \quad x^{(j_{p_k})} - c_k \Big], \tag{15}$$

and the centered feature matrix by $X_w = [X_{1,c}, \ldots, X_{K,c}]$. The intra-class spread matrix is then defined by $S_w := X_w X_w^\top$. We now focus on defining inter-class spread. Let us define the global centroid of the data by

$$c := \frac{1}{p} \sum_{i=1}^{p} x^{(i)}. \tag{16}$$

We would like to measure the distance of the centroid of each cluster to the global centroid. Therefore we define the following matrix

$$\bar{X} := [\underbrace{(c_1 - c), \ldots, (c_1 - c)}_{p_1 \text{ times}}, \ldots, \underbrace{(c_k - c), \ldots, (c_k - c)}_{p_k \text{ times}}] \in \mathbb{R}^{n \times p}, \tag{17}$$

whose columns measures the distance of each centroid to the global centroid. Finally the inter-class spread matrix can be defined by

$$S_b := \bar{X}\bar{X}^\top. \tag{18}$$

As we discussed earlier in this section, we would like to maximize inter-class spread while minimizing the intra-class spread. To do so we can scalarize the above two metrics by defining the following scalar function:

$$H(\mathsf{q}) := \frac{\mathsf{q}^\top S_b \mathsf{q}}{\mathsf{q}^\top S_w \mathsf{q}}, \quad \forall q \in \mathbb{R}^n, q \neq 0. \tag{19}$$

Any direction q that maximizes $H(\mathsf{q})$ can be viewed as a desirable direction maximizes the ratio of the spread of inter-class over the spread of the intra-class. Note that such direction does not necessarily maximizes spread of inter-class and minimizes spread of intra-class simultaneously.

Let us suppose for the moment that S_w is symmetric positive definite (s.p.d.). Then using Rayleigh quotient of Section 1.1 we can conclude that the maximum value of $H(\mathsf{q})$ corresponds to the maximum eigenvalue of $S_w^{-1}S_b$ and the optimum direction is the corresponding eigenvector of $S_w^{-1}S_b$, i.e.,

$$v_{\max} = \arg\max_{\mathsf{q} \in \mathbb{R}^n} H(\mathsf{q}), \tag{20}$$

where

$$S_w^{-1} S_b \, v_{\max} = \lambda_{\max} \, v_{\max}. \tag{21}$$

For the case where S_w is only semi-definite, we cannot use the argument that S_w^{-1} exists. One approach would be to regularize S_w to an s.p.d. matrix by adding an s.p.d. matrix, e.g.,

$$S_{w,\varepsilon} := S_w + \varepsilon I, \quad \varepsilon > 0. \tag{22}$$

Here ε is a small positive real number.

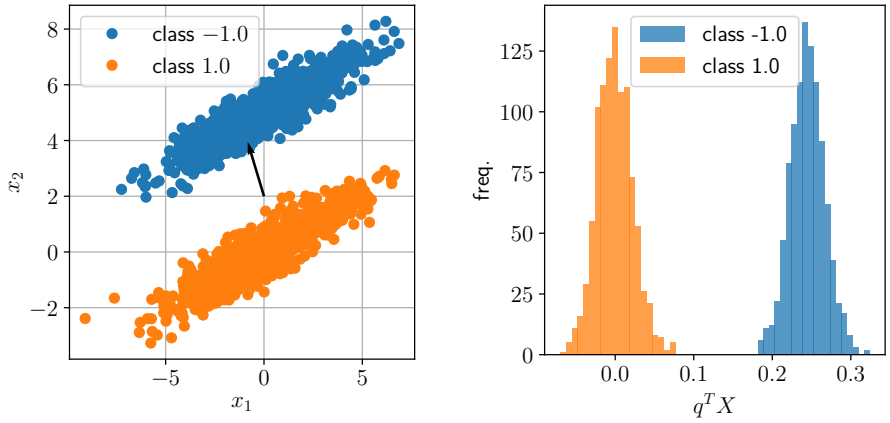

Figure 1: Synthetic dataset from two Gaussian distributions with the LDA direction q as black arrow (left), and projected dataset into the subspace defined by LDA (right).

3.1.2 Numerical example

In this section, we would like to demonstrate the effectiveness of LDA on an annotate data set. We consider a synthetic dataset originating from two Gaussian distribution (corresponding to two distinct classes, i.e., $K = 2$) with the probability density functions

$$f_X(x) \propto e^{-\frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k)}, \quad k = 1, 2, \tag{23}$$

with covariance matrix

$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} 5 & 2 \\ 2 & 1 \end{bmatrix}, \tag{24}$$

and means

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 0 \\ 10 \end{pmatrix}. \tag{25}$$

For a realization of this dataset see Figure 1 (left). In Figure 1 (right) we find the projection of the dataset in the subspace defined by q. Note that the two classes are well-separated in this subspace.

3.2 Logistic regression

Logistic regression is considered as a simple method for the classification problem. It is considered as a discriminative classification method since it models the posterior probability of the target variable directly. Let us denote as before the dataset by the tuple (X, y) where $X \in \mathbb{R}^{n\times p}$ is the so-called feature matrix, and $y \in \mathbb{B}^p$ is the target vector. Here p denotes the number of data points in the dataset and n is the dimension of the feature space. Logistic regression models the posterior probability of the target variable directly through a sigmoid function, i.e.,

$$\mathbb{P}(y = 1|x) = \sigma(\mathbf{w}^\top x + w_0), \quad \mathbf{w} \in \mathbb{R}^n, w_0 \in \mathbb{R}, \tag{26}$$

and from the definition of the probabilities we have that $\mathbb{P}(y = 0|x) = 1 - \mathbb{P}(y = 1|x)$.

Here $\sigma(a) := 1/(1 + \exp(-a))$. Note that when $\mathbf{w}^\top x + w_0 = 0$ then $\mathbb{P}(y = 0|x) = \mathbb{P}(y = 1|x) = \frac{1}{2}$ and we can conclude that the logistic regression model does not prefer one class over the other given the input x. Such a $(n - 1)$-dimensional hyperplane is called the decision surface. Logistic regression is called a linear classification method since its decision boundary is a linear function of the input x.

In order to find optimal parameters of logistic regression model, i.e., $\mathbf{w}$ and w_0, we first form a log-likelihood function over the posterior distribution and then maximize it. More precisely for a given logistic model (fixed $\mathbf{w}$ and w_0) the probability of observing $(y^{(j)}, x^{(j)})$ is given by

$$\mathbb{P}\big((y^{(j)}, x^{(j)})|\mathbf{w}, w_0\big) = \mathbb{P}(y^{(j)} = 1|x^{(j)})^{y^{(j)}} \cdot \mathbb{P}(y^{(j)} = 0|x^{(j)})^{(1-y^{(j)})}. \tag{27}$$

Since observations of the dataset are i.i.d. we can conclude that the probability of observing the dataset given the model is the multiplication of the above likelihood function, i.e.,

$$\mathbb{P}\big((y, X)|\mathbf{w}, w_0\big) = \prod_{j=1}^{p} \mathbb{P}\big((y^{(j)}, x^{(j)})|\mathbf{w}, w_0\big). \tag{28}$$

4 Kernel methods

In this section we will go through a class of methods known as *kernel methods* for regression and classification. Let us start from a regression model defined by

$$f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^{d_0}, \tag{29}$$

where $\mathbf{x}$ is the original feature inputs which are transformed using a feature map $\phi : \mathbb{R}^{d_0} \to \mathbb{R}^d$ and $\mathbf{w} \in \mathbb{R}^d$ is the weight that completes the definition of the method. We refer to $\mathbb{R}^d$ as the feature space while $\mathbb{R}^{d_0}$ is the initial input space. In this section we assume that the feature space has a much higher dimensionality compare to initial feature space, i.e., $d \gg d_0$.

Example 4.1 *Suppose that we have an image of size* 16×16 *pixels where each pixel's intensity is represented by an entry in a vector denoted by* $\boldsymbol{x} \in \mathbb{R}^{d_0}$*. Suppose we would like to construct another vector consisting of all monomials of degree upto* m*, e.g., for* $m = 2$*, we have*

$$\phi(\boldsymbol{x}) = (x_1, \ldots, x_{d_0}, x_1x_2, x_1x_3, \ldots, x_1^2, \ldots, x_{d_0}^2)^\top \in \mathbb{R}^d, \tag{30}$$

where $d = \begin{pmatrix} 2 + d_0 - 1 \\ d_0 - 1 \end{pmatrix} = O(d_0^2)$*. In general, if we are interested in arbitrary* $m \in \mathbb{N}^+$*, we have* $d = O\big(\frac{d_0^m}{m!}\big)$*.*

Suppose we have a set of data points $\{(\mathbf{x}_i, t_i)\}_{i=1}^N$ which we can use to fit weights, i.e., $\mathbf{w}$. In order to do so, let us setup a loss (objective) function and an optimization, i.e.,

$$J(\mathbf{w}) := \frac{1}{2}\sum_{i=1}^{N}(t_i - \mathbf{w}^\top\phi(\mathbf{x}_i))^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2, \tag{31}$$

and find $\mathbf{w}^\star$ such that

$$\mathbf{w}^\star = \arg\min_{\mathbf{w}\in\mathbb{R}^d} J(\mathbf{w}). \tag{32}$$

In order to solve for the optimal weights we need to take the derivative of $J(\mathbf{w})$ with respect to its argument and set it to zero:

$$\nabla J(\mathbf{w}) = \lambda\mathbf{w} - \sum_{i=1}^{N}(t_i - \mathbf{w}^\top\phi(\mathbf{x}_i))\phi(\mathbf{x}_i) = 0. \tag{33}$$

In order to find the solution to this equation, we first define $X := \big[\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_N)\big]$ and $\mathbf{t} := (t_1, \ldots, t_N)^\top$. Then the optimal $\mathbf{w}^\star$ reads

$$\big(\lambda I_d + XX^\top\big)\mathbf{w}^\star = X\mathbf{t}. \tag{34}$$

On the other hand, let us rewrite the optimal weights in the following form

$$\mathbf{w}^\star = \frac{1}{\lambda}X\mathbf{t} - XX^\top\mathbf{w}^\top = X\Big(\frac{1}{\lambda}\mathbf{t} - X^\top\mathbf{w}^\star\Big) =: X\mathbf{a}, \tag{35}$$

where $\mathbf{a} \in \mathbb{R}^N$. Substituting $\mathbf{w}^\star = X\mathbf{a}$ into the function definition yields

$$f^\star(\mathbf{x}) = \mathbf{a}^\top X^\top \phi(x) = \sum_{j=1}^{N} a_i \phi(\mathbf{x}_i)^\top \phi(x) \quad \forall \mathbf{x} \in \mathbb{R}^{d_0}. \tag{36}$$

This implies that although we search in a rich function space for the optimal weights, i.e., $\mathbb{R}^d$, the optimal solution lies in a much smaller space, namely the space spanned by the training dataset $\{\phi(\mathbf{x}_i)\}_{i=1}^N$. Often the problem is posed in a setting where $N \ll d$, e.g., see Example 4.1. Therefore it might be much cheaper to pose the original function space and optimization in terms of $\mathbf{a}$. We call this representation of the problem the *dual representation*.

The dual representation in terms of $\mathbf{a}$ is the following:

$$J(\mathbf{a}) := \frac{1}{2} \sum_{i=1}^{N} (t_i - \mathbf{a}^\top X^\top \phi(\mathbf{x}_i))^2 + \frac{\lambda}{2} \mathbf{a}^\top X^\top X \mathbf{a}, \tag{37}$$

and $\mathbf{a}^\star := \arg\min_{\mathbf{a} \in \mathbb{R}^N} J(\mathbf{a})$. The optimal $\mathbf{a}^\star$ satisfies

$$(\lambda I_N + X^\top X)\mathbf{a}^\star = \mathbf{t}. \tag{38}$$

The solution of the primal and dual representations coincide in the sense that

$$\mathbf{w}^\star = X\mathbf{a}^\star. \tag{39}$$

Remark 4.1 *The two function spaces of the primal and dual representations do not coincide. Observe that $\mathbf{a} \in \mathbb{R}^N$ while $\boldsymbol{w} \in \mathbb{R}^d$ and $N \ll d$. Therefore the primal representation has a richer function space, however the optimal solution of both formulations lies in $\mathbb{R}^N$.*

4.1 Kernel trick

Recall that in the dual representation the regression function is

$$f_{\mathbf{a}}(\mathbf{x}) = \sum_{j=1}^{N} a_j \phi(\mathbf{x}_j)^\top \phi(\mathbf{x}). \tag{40}$$

Note that we can replace $\phi(\mathbf{x}_i)^\top \phi(\mathbf{x})$ by a *kernel* function $k(\mathbf{x}_i, \mathbf{x}) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x})$ in order to avoid carrying around the high dimensional map $\phi(\mathbf{x})$. More precisely, we need to find a kernel function $k(\mathbf{x}, \mathbf{y})$ that admits a decomposition $k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^\top \phi(\mathbf{y})$ for some $\phi(\cdot)$. A necessary and sufficient condition for a valid kernel function is that the matrix K, whose entries are $K_{mn} = k(\mathbf{x}_m, \mathbf{x}_n)$ for all sets of $\{\mathbf{x}_i\}_{i=1}^N$, is positive semi-definite. The matrix K is called *Gram matrix*.

Example 4.2 *The kernel corresponding to the problem posed in Example 4.1 is* $k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^\top \boldsymbol{x}')^2$ *for all* $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{d_0}$.