SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Intro to Big Data Science: Assignment 1

Due Date: Mar 15, 2024

✏ Exercise 1

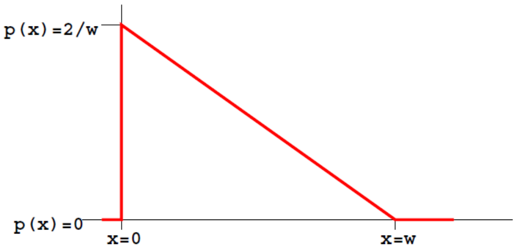
Given the ordered data $\{x_{(i)}\}_{i=1}^{2n-1}$ with increasing order. Show that the median of the data set is equal to the minimizer of the following L^1 minimization problem:

$$x_{(n)} = \arg\min_{c} \sum_{i=1}^{2n-1} |x_{(i)} - c|.$$

✏ Exercise 2

Consider the probability density function (PDF) shown in the following figure and equations:

$$p(x) = \begin{cases} 0, & \text{if} x < 0, \\ \frac{2}{w} - \frac{2x}{w^2}, & \text{if} 0 \leqslant x \leqslant w, \\ 0, & \text{if} w < x. \end{cases}$$



1. Which of the following expression is true? (Only one truth.)

(A) $E[X] = \int_{-\infty}^{\infty} (\frac{2}{w} - \frac{2x}{w^2}) \mathrm{d}x$;

(B) $E[X] = \int_{-\infty}^{\infty} x(\frac{2}{w} - \frac{2x}{w^2}) \mathrm{d}x$;

(C) $E[X] = \int_{-\infty}^{\infty} w(\frac{2}{w} - \frac{2x}{w^2}) \mathrm{d}x$;

(D) $E[X] = \int_{0}^{w} (\frac{2}{w} - \frac{2x}{w^2}) \mathrm{d}x$;

(E) $E[X] = \int_{0}^{w} x(\frac{2}{w} - \frac{2x}{w^2}) \mathrm{d}x$;

(F) $E[X] = \int_{0}^{w} w(\frac{2}{w} - \frac{2x}{w^2}) \mathrm{d}x$;

2. What is $\mathbb{P}(x = 1 | w = 2)$?
3. When $w = 2$, what is $p(1)$?

Exercise 3

Let X and Y be two continuous random variables. The conditional expectation of Y on $X = x$ is defined as the expectation of Y with respect to the conditional probability density $p(Y|X)$:

$$\mathrm{E}(Y|X = x) = \int_{\mathscr{Y}} y p(y|X = x) \mathrm{d}y = \frac{\int_{\mathscr{Y}} y p(x, y) \mathrm{d}y}{p_x(x)},$$

where $p_x(x)$ is the marginal probability density of Y. Show the following properties of the conditional expectation:

1. $\mathrm{E}_{p_y} Y = \mathrm{E}_{p_x}[\mathrm{E}(Y|X)]$, where E_{p_y} means taking the expectation with respect to the marginal probability density p_y.
   Remark: This formula is sometimes called the tower rule.
2. If X and Y are independent, then $\mathrm{E}(Y|X = x) = \mathrm{E}(Y)$.
3. The minimizer of the following minimization problem with respect to some constant $c \in \mathbb{R}$
   $$\arg\min_{c} \mathrm{E}[(Y - c)^2 | X = x]$$
   is attained at $c^* = \mathrm{E}(Y|X = x)$.

Exercise 4 Hereby, I would first say sorry for my mistake made in class: "the rand distance does not satisfy the triangle inequality" is wrong. Here in this problem you can show it.

1. The symmetric distance (or rand distance) between two sets $A \subset \Omega$ and $B \subset \Omega$ is defined as $R_\delta(A, B) = \frac{|A \backslash B| + |B \backslash A|}{|\Omega|} = \frac{|A \triangle B|}{|\Omega|}$, where $|S|$ stands for the number of elements in the set S. Show that the rand distance R_δ is actually a distance, i.e., it satisfies the three properties:

a) Positivity: $R_\delta(A,B) \geqslant 0$, and "=" if and only if $A = B$;

b) Symmetry: $R_\delta(A,B) = R_\delta(B,A)$;

c) Triangle inequality: $R_\delta(A,B) \leqslant R_\delta(A,C) + R_\delta(B,C)$.

2. (Optional) The Jaccard distance between two sets A and B is defined as $J_\delta(A,B) = 1 - \frac{|A \bigcap B|}{|A \bigcup B|} = \frac{|A \triangle B|}{|A \bigcup B|}$. Show that the Jaccard distance J_δ is also a distance satisfying the three above properties.

SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Intro to Big Data Science: Assignment 2

Due Date: March 29, 2024

✏ **Exercise 1**

Log into “cookdata.cn”, and enroll the course “数据科学导引”. Finish the online exercise there.

✏ **Exercise 2 (Decision Tree)**

You are trying to determine whether a boy finds a particular type of food appealing based on the food’s temperature, taste, and size.

Food Sample Id	Appealing	Temperature	Taste	Size
1	No	Hot	Salty	Small
2	No	Cold	Sweet	Large
3	No	Cold	Sweet	Large
4	Yes	Cold	Sour	Small
5	Yes	Hot	Sour	Small
6	No	Hot	Salty	Large
7	Yes	Hot	Sour	Large
8	Yes	Cold	Sweet	Small
9	Yes	Cold	Sweet	Small
10	No	Hot	Salty	Large

1. What is the initial entropy of “Appealing”?
2. Assume that “Taste” is chosen as the root of the decision tree. What is the information gain associated with this attribute.
3. Draw the full decision tree learned from this data (without any pruning).

✏ **Exercise 3: (Maximum Likelihood Estimate (MLE, 极大似然估计))**

Suppose that the samples $\{x_i\}_{i=1}^n$ are drawn from Normal distribution $\mathcal{N}(\mu,\sigma^2)$ with p.d.f. $f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$, where $\theta = (\mu,\sigma^2)$. The Maximum likelihood estimator (MLE) of θ is the one that maximize the likelihood function

$$L(\theta) = \prod_{i=1}^n f_\theta(x_i)$$

1. Show that the MLE estimator of the parameters (μ,σ^2) is

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2.$$

2. Show that

$$\mathrm{E}\hat{\mu} = \mu, \qquad \mathrm{E}\Big(\frac{n}{n-1}\hat{\sigma}^2\Big) = \sigma^2,$$

where E is the expectation. This means that $\hat{\mu}$ is an unbiased estimator of μ, but $\hat{\sigma}^2$ is a biased estimator of σ^2.

✏ **Exercise 4 (MLE for Naive Bayes methods)**

Suppose that X and Y are a pair of discrete random variables, i.e., $X \in \{1,2,\ldots,t\}$, $y \in \{1,2,\ldots,c\}$. Then the probability distribution of Y is solely dependent on the set of parameters $\{p_k\}_{k=1}^c$, where $p_k = \Pr(Y = k)$ with $\sum_{k=1}^c p_k = 1$. Similarly, the conditional probability distribution of X given Y is solely dependent on the set of parameters $\{p_{sk}\}_{k=1,\ldots,c}^{s=1,\ldots,t}$, where $p_{sk} = \Pr(X = s|Y = k)$ with $\sum_{s=1}^t p_{sk} = 1$. Now we have a set of samples $\{(x_i,y_i)\}_{i=1}^n$ drawn independently from the joint distribution $\Pr(X,Y)$. Prove that the MLE of the parameter p_k (prior probability) is

$$\hat{p}_k = \frac{\sum_{i=1}^n \mathrm{I}(y_i = k)}{n}, k = 1,\ldots,c;$$

and the MLE of the parameter p_{ks} is

$$\hat{p}_{sk} = \frac{\sum_{i=1}^n \mathrm{I}(x_i = s, y_i = k)}{\sum_{i=1}^n \mathrm{I}(y_i = k)}, s = 1,\ldots,t, k = 1,\ldots,c.$$

✏ **Exercise 5 (Error bound for 1-nearest-neighbor method, optional)** In class, we have estimated that the error for 1-nearest-neighbor rule is roughly twice the Bayes error. Now let us make it more rigorous.

Let us consider the two-class classification problem with $\mathscr{X} = [0,1]^d$ and $\mathscr{Y} = \{0,1\}$. The underlying joint probability distribution on $\mathscr{X} \times \mathscr{Y}$ is $P(\mathbf{X}, Y)$ from which we deduce that the marginal distribution of $\mathbf{X}$ is $p_{\mathbf{X}}(\mathbf{x})$ and the conditional probability distribution is $\eta(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x})$. Assume that $\eta(\mathbf{x})$ is c-Lipschitz continuous: $|\eta(\mathbf{x}) - \eta(\mathbf{x}')| \leqslant c\|\mathbf{x} - \mathbf{x}'\|$ for any $\mathbf{x},\mathbf{x}' \in \mathscr{X}$. Recall that the Bayes rule is $f^*(\mathbf{x}) = 1_{\{\eta(\mathbf{x})>1/2\}}$. Given a training set

$S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $(\mathbf{x}_i, y_i) \overset{i.i.d.}{\sim} P$ (or equivalently $S \sim P^n$), the 1-nearest-neighbor rule is $f^{1NN}(\mathbf{x}) = y_{\pi_S(\mathbf{x})}$ where $\pi_S(\mathbf{x}) = \arg\min_i \|\mathbf{x} - \mathbf{x}_i\|$.

Define the generalization error for rule f as $\mathscr{E}(f) = \mathrm{E}_{(\mathbf{X},Y)\sim P} 1_{Y \neq f(\mathbf{X})}$. Show that

$$\mathrm{E}_{S\sim P^n} \mathscr{E}(f^{1NN}) \leqslant 2\mathscr{E}(f^*) + c\mathrm{E}_{S\sim P^n} \mathrm{E}_{\mathbf{x}\sim p_{\mathbf{X}}} \|\mathbf{x} - \mathbf{x}_{\pi_S(\mathbf{x})}\|.$$

(This means that we can have a precise error estimate for 1-nearest-neighbor rule if we can bound $\mathrm{E}_{S\sim P^n} \mathrm{E}_{\mathbf{x}\sim p_{\mathbf{X}}} \|\mathbf{x} - \mathbf{x}_{\pi(\mathbf{x})}\|$.)

South University of Science and Technology of China

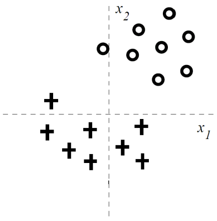
Intro to Big Data Science: Assignment 3

Due Date: April 19, 2024

Exercise 1

Log into Cookdata, and enroll the course “数据科学导引”. Finish the online exercise there.

Exercise 2 We consider here a discriminative approach for solving the classification problem illustrated in the following figure, where “+” corresponds to class $y = 1$ and “O” corresponds to class $y = 0$:



1. We attempt to solve the binary classification task depicted in Figure 1 with the simple linear logistic regression model

$$P(y=1|\mathbf{x},\mathbf{w}) = \sigma(w_0 + w_1x_1 + w_2x_2) := \frac{1}{1+\exp(-w_0 - w_1x_1 - w_2x_2)}.$$

where $\mathbf{x} = (x_1, x_2)^T$, $\mathbf{w} = (w_0, w_1, w_2)^T$, and σ is the sigmoid function defined by the last equality. Notice that the training data can be separated with zero training error with a linear separator.

Consider training regularized linear logistic regression models for the training set $\{(\mathbf{x}^{(i)}, y^{(i)})\}$, where we try to maximize

$$\sum_{i=1}^{n} \log P(y^{(i)}|\mathbf{x}^{(i)}, \mathbf{w}) - Cw_j^2,$$

for very large C. The regularization penalties used in penalized conditional log-likelihood estimation are $-Cw_j^2$ where $j = \{0, 1, 2\}$. In other words, only one of the parameters is regularized in each case. Given the training data in the above figure, how does the training error changes with regularization of each parameter w_j? State whether the training error increases or stays the same (zero) for each w_j $(j = 0, 1, 2)$ for very large C. Provide a brief justification for each of your answers.

2. If we change the form of regularization to L_1-norm (absolute value) and regularize w_1 and w_2 only (but not w_0), we get the following penalized log-likelihood

$$\sum_{i=1}^{n} \log P(y^{(i)}|\mathbf{x}^{(i)}, \mathbf{w}) - C(|w_1| + |w_2|).$$

Consider again the problem in the above figure and the same linear logistic regression model $P(y = 1|\mathbf{x}, \mathbf{w}) = \sigma(w_0 + w_1 x_1 + w_2 x_2)$.

a) As we increase the regularization parameter C, which of the following scenarios do you expect to observe? (Choose only one) Briefly explain your choice.

(A) First w_1 will become 0, then w_2.

(B) First w_2 will become 0, then w_1.

(C) w_1 and w_2 will become zero simultaneously.

(D) None of the weights will become exactly zero, only smaller as C increases.

b) For very large C, with the same L_1-norm regularization for w_1 and w_2 as above, which value(s) do you expect w_0 to take? Explain your answer by doing some calculation. (Note that the number of points from each class is the same.) (You can give a range of values for w_0 if you deem necessary).

c) Assume that we obtain more data points from the "+" class that corresponds to $y = 1$ so that the class labels become unbalanced. Again for very large C, with the same L_1-norm regularization for w_1 and w_2 as above, which value(s) do you expect w_0 to take? Explain briefly. (You can give a range of values for w_0 if you deem necessary).

✏ Exercise 3 (Linear regression)

Consider a multivariate liner model $\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$ with $\mathbf{y} \in \mathbb{R}^{n\times 1}$, $\mathbf{X} \in \mathbb{R}^{n\times(d+1)}$, $\mathbf{w} \in \mathbb{R}^{(d+1)\times 1}$, and $\epsilon \in \mathbb{R}^{n\times 1}$, where $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, follows the normal distribution.

1. Show that the linear regression predictor is given by $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

2. Let $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, show that $\mathbf{P}$ has only 0 and 1 eigenvalues.

3. Show that $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is an unbiased estimator of $\mathbf{w}$, i.e., $\mathbf{E}(\hat{\mathbf{w}}) = \mathbf{w}$. Also show that $\mathrm{Var}(\hat{\mathbf{w}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$. (Note that by definition, $\mathrm{Var}(\hat{\mathbf{w}}) = \mathbf{E}[(\hat{\mathbf{w}} - \mathbf{E}(\hat{\mathbf{w}}))(\hat{\mathbf{w}} - \mathbf{E}(\hat{\mathbf{w}}))^T]$).
4. Recall the definition of R^2 score: $R^2 := 1 - \frac{SS_{res}}{SS_{tot}}$, where $SS_{tot} = \sum_{i=1}^n (y_i - \bar{y})^2$, $SS_{reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, and $SS_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$. Prove that for linear regression, $SS_{tot} = SS_{reg} + SS_{res}$. (So that R^2 score can also be defined as $R^2 = \frac{SS_{reg}}{SS_{tot}}$)
5. Repeat the questions in 1 and 3 if we are using ridge regression with a regularization parameter λ.

✏ **Exercise 4 (Generalized Cross-Validation)** Consider ridge regression:

$$\min_{\mathbf{w}} \left[(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2 \right]$$

It has the solution $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$ and prediction $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{P}\mathbf{y}$ with $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T$ be the projection matrix.

1. Define the leave-one-out cross validation estimator as

   $$\hat{\mathbf{w}}^{[k]} = \arg\min_{\mathbf{w}} \left[\sum_{i=1, i\neq k}^{n} (y_i - \mathbf{x}_i^T\mathbf{w})^2 + \lambda\|\mathbf{w}\|_2^2 \right].$$

   Show that $\hat{\mathbf{w}}^{[k]} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} - \mathbf{x}_k\mathbf{x}_k^T)^{-1}(\mathbf{X}^T\mathbf{y} - \mathbf{x}_k y_k)$
2. (Optional) Define the ordinary cross-validation (OCV) mean squared error as $V_0(\lambda) = \frac{1}{n}\sum\limits_{k=1}^{n}(\mathbf{x}_k^T\hat{\mathbf{w}}^{[k]} - y_k)^2$. Show that $V_0(\lambda)$ can be rewritten as $V_0(\lambda) = \frac{1}{n}\sum\limits_{k=1}^{n}\left(\frac{\hat{y}_k - y_k}{1 - p_{kk}}\right)^2$, where $\hat{y}_k = \sum_{j=1}^n p_{kj} y_j$ and p_{kj} is the (k, j)-entry of $\mathbf{P}$.

   (Hint: You may need to use the Sherman-Morrison Formula for nonsingualar matrix $\mathbf{A}$ and vectors $\mathbf{x}$ and $\mathbf{y}$ with $\mathbf{y}^T\mathbf{A}^{-1}\mathbf{x} \neq -1$: $(\mathbf{A} + \mathbf{x}\mathbf{y}^T)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{x}\mathbf{y}^T\mathbf{A}^{-1}}{1 + \mathbf{y}^T\mathbf{A}^{-1}\mathbf{x}}\Big)$
3. (Optional) Define weights as $w_k = \left(\frac{1 - p_{kk}}{\frac{1}{n}tr(\mathbf{I} - \mathbf{P})}\right)^2$ and weighted OCV as $V(\lambda) = \frac{1}{n}\sum\limits_{k=1}^{n} w_k(\mathbf{x}_k^T\hat{\mathbf{w}}^{[k]} - y_k)^2$. Show that $V(\lambda)$ can be written as

   $$V(\lambda) = \frac{\frac{1}{n}\|(\mathbf{I} - \mathbf{A})\mathbf{y}\|^2}{\left[1 - tr(\mathbf{P})/n\right]^2}$$

✏ **Exercise 5** (Solving LASSO by ADMM) The alternating direction method of multipliers (ADMM) is a very useful algorithm for solving the constrained optimization problem:

$$\min_{\theta, z} f(\boldsymbol{\theta}) + g(\mathbf{z}), \qquad \text{subject to} \quad \mathbf{A}\boldsymbol{\theta} + \mathbf{B}\mathbf{z} = \mathbf{c}.$$

The algorithm is given by using Lagrange multiplier $\mathbf{u}$ for the constraint. The detail is as follows:

1. $\boldsymbol{\theta}^{(k+1)} = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathbf{z}^{(k)}, \mathbf{u}^{(k)})$;

2. $\mathbf{z}^{(k+1)} = \arg\min_{\mathbf{z}} L(\boldsymbol{\theta}^{(k+1)}, \mathbf{z}, \mathbf{u}^{(k)})$;

3. $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \mathbf{A}\boldsymbol{\theta}^{(k+1)} + \mathbf{B}\mathbf{z}^{(k+1)} - \mathbf{c}$;

where L is the augmented Lagrange function defined as

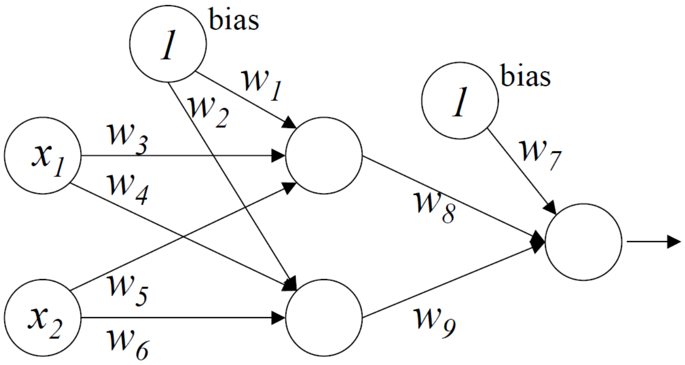
$$L(\boldsymbol{\theta}, \mathbf{z}, \mathbf{u}) = f(\boldsymbol{\theta}) + g(\mathbf{z}) + \mathbf{u}^T(\mathbf{A}\boldsymbol{\theta} + \mathbf{B}\mathbf{z} - \mathbf{c}) + \frac{1}{2}\|\mathbf{A}\boldsymbol{\theta} + \mathbf{B}\mathbf{z} - \mathbf{c}\|_2^2.$$

An advantage of ADMM is that no tuning parameter such as the step size in the gradient algorithm is involved. Please write down the ADMM steps for solving LASSO problem:

$$\min_{\mathbf{w}} \left[\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1 \right].$$

(Hint: In order to use ADMM, you have to introduce an auxiliary variable and a suitable constraint. Please give the explicit formulae by solving "arg min" in each step of ADMM.)

✏ **Problem 6** (NeuralNet) Consider a neural network for a binary classification which has one hidden layer as shown in the figure. We use a linear activation function $h(z) = cz$ at hidden units and a sigmoid activation function $g(z) = \frac{1}{1+\exp(-z)}$ at the output unit to learn the function for $\mathrm{P}(y = 1|\mathbf{x}, \mathbf{w})$ where $\mathbf{x} = (x_1, x_2)$ and $\mathbf{w} = (w_1, \ldots, w_9)$.



1. What is the output $\mathrm{P}(y = 1|\mathbf{x}, \mathbf{w})$ from the above neural net? Express it in terms of x_i, c and weights w_i. What is the final classification boundary?

2. Draw a neural net with no hidden layer which is equivalent to the given neural net, and write weights $\tilde{\mathbf{w}}$ of this new neural net in terms of c and w_i.

3. Is it true that any multi-layered neural net with linear activation functions at hidden layers can be represented as a neural net without any hidden layer? Briefly explain your answer.

SOUTH UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA

Intro to Big Data Science: Assignment 4

Due Date: May 10, 2024

Exercise 1

Log into “cookdata.cn”, and enroll the course “数据科学导引”. Finish the online exercise there.

Problem 2 (Support Vector Machine (SVM)) Soft-Margin Linear SVM. Given the following dataset aligning on the x-axis (See the figure below), which consists of 4 positive data points $\{0,1,2,3\}$ and 3 negative data points $\{-3,-2,-1\}$. Suppose that we want to learn a soft-margin linear SVM for this data set. Remember that the soft-margin linear SVM can be formalized as the following constrained quadratic optimization problem. In this formulation, C is the regularization parameter, which balances the size of margin (i.e., smaller $\|\mathbf{w}\|_2^2$) vs. the violation of the margin (i.e., smaller $\sum_{i=1}^{m}\xi_i$).

$$
\begin{aligned}
&\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i \\
&\text{s.t.}\ \ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geqslant 1-\xi_i, \quad \xi_i \geqslant 0, \quad i = 1,\dots,n
\end{aligned}
$$

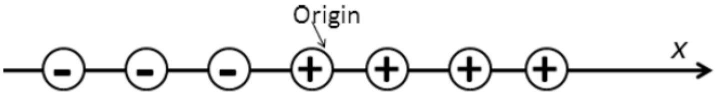
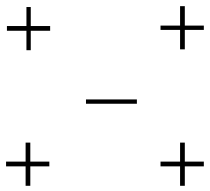


Figure 1: The data set.

1. If $C = 0$, which means that we only care the size of the margin, how many support vectors do we have?

2. if $C \to \infty$, which means that we only care the violation of the margin, how many support vectors do we have?
3. Properties of Kernel:
   a) Using the definition of kernel functions in SVM, prove that the kernel $K(\mathbf{x}_i, \mathbf{x}_j)$ is symmetric, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the feature vectors for i-th and j-th examples.
   b) Given n training examples $(\mathbf{x}_i, \mathbf{x}_j)$ for $(i, j = 1, \ldots, n)$, the kernel matrix A is an $n \times n$ square matrix, where $A(i, j) = K(\mathbf{x}_i, \mathbf{x}_j)$. Prove that the kernel matrix A is semi-positive definite.

Exercise 3 Consider training an AdaBoost classifier using decision stumps on the five-point data set (4 "+" samples and 1 "-" sample):



1. Which examples will have their weights increased at the end of the first iteration? Circle them.
2. How many iterations will it take to achieve zero training error? Explain by doing some computation using the above algorithm.
3. Can you add one more sample to the training set so that AdaBoost will achieve zero training error in two steps? If not, explain why.

Exercise 4 (Hierarchical Clustering)

Suppose that we have four observations, for which we compute a dissimilarity matrix, given by

$$\begin{pmatrix} 0 & 0.3 & 0.4 & 0.7 \\ 0.3 & 0 & 0.5 & 0.8 \\ 0.4 & 0.5 & 0 & 0.45 \\ 0.7 & 0.8 & 0.45 & 0 \end{pmatrix}$$

For instance, the dissimilarity between the first and second observations is 0.3, and the dissimilarity between the second and fourth observations is 0.8.

1. On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observations using complete linkage. Be sure to indicate on the plot the height at which each fusion (merge) occurs, as well as the observations corresponding to each leaf in the dendrogram.
2. Repeat 1, this time using single linkage clustering.

3. Suppose that we cut the dendrogram obtained in 1 such that two clusters result. Which observations are in each cluster?

4. Suppose that we cut the dendrogram obtained in 2 such that two clusters result. Which observations are in each cluster?

5. It is mentioned in the chapter that at each fusion in the dendrogram, the position of the two clusters being fused can be swapped without changing the meaning of the dendrogram. Draw a dendrogram that is equivalent to the dendrogram in 1, for which two or more of the leaves are repositioned, but for which the meaning of the dendrogram is the same.

✏ **Exercise 5** In this problem, you need to show that the within-cluster point scatter (or in other words, the sum-of-squared errors (SSE)) is **non-increasing** when the number of clusters increases.

Consider a data set $\mathscr{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ that contains N observations. Each sample $\mathbf{x}_i$ is a d-dimensional vector of continuous-valued attributes. You are performing K-means clustering.

1. Suppose all the N samples are grouped into a **single** cluster. Let μ be the centroid of the cluster. Express the total sum-of-squared errors SSE_T in terms of $\mathbf{x}_i$, μ and N. And show that SSE_T can be decomposed into d separate terms, one for each attribute, i.e., $SSE_T = \sum_{j=1}^{d} SSE_j$.

2. Now, suppose all the N observations are grouped into two clusters, C_1 and C_2. Let μ_1 and μ_2 be their corresponding cluster centroids while n_1 and n_2 are their respective cluster sizes ($n_1 + n_2 = N$). Express the sum-of-squared errors for each cluster, $SSE^{(j)}$ ($j = 1$ or 2), in terms of $\mathbf{x}_i$, n_j, and μ_j. You need to expand the quadratic term, $(a-b)^2 = a^2 - 2ab + b^2$, and simplify the expression.

3. By rewriting your expression for SSE_T in terms of $\mathbf{x}_i$, n_1, n_2, μ_1, μ_2 and N, show that $SSE_T \geqslant SSE^{(1)} + SSE^{(2)}$.

SOUTH UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA

Intro to Big Data Science: Assignment 5

Due Date: May 31, 2024

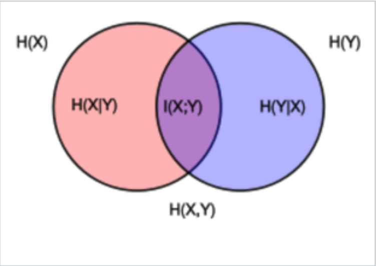
✏ **Exercise 1**

Log into “cookdata.cn”, and enroll the course “数据科学导引”. Finish the online exercise there.

✏ **Exercise 2** Recall the definition of information entropy, $H(P) = -\sum_{i=1}^{n} p_i \log p_i$, which means the maximal information contained in probability distribution P. Let X and Y be two random variables. The entropy $H(X,Y)$ for the joint distribution of (X,Y) is defined similarly. The conditional entropy is defined as:

$$\begin{aligned} H(X|Y) &= -\sum_j P(Y=y_j) H(X|Y=y_j) \\ &= -\sum_j P(Y=y_j) (\sum_i P(X=x_i|Y=y_j) \log P(X=x_i|Y=y_j)) \end{aligned}$$

1. Show that $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$.
2. The mutual information (information gain) is defined as $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$. Show that if X and Y are independent, then $I(X;Y) = 0$



3. Define the Kullback-Leibler divergence as $D_{KL}(P\|Q) = -\sum_{i=1}^{n} p_i \log \frac{q_i}{p_i}$. Show that $I(X;Y) = D_{KL}(p(X,Y)\|p(X)p(Y))$.
4. (Optional) Furthermore, show that $D_{KL}(P\|Q) \geqslant 0$ for any P and Q by using Jensen's inequality. As a result, $I(X;Y) \geqslant 0$.

Exercise 3 (EM Algorithm, you may need to carefully read Section 8.5.2 in the book "Elements of Statistical Learning" before solving this problem)

Imagine a class where the probability that a student gets an "A" grade is $\mathbb{P}(A) = \frac{1}{2}$, a "B" grade is $\mathbb{P}(B) = \mu$, a "C" grade is $\mathbb{P}(C) = 2\mu$, and a "D" grade is $\mathbb{P}(D) = \frac{1}{2} - 3\mu$. We are told that c students get a "C" and d students get a "D". We don' t know how many students got exactly an "A" or exactly a "B". But we do know that h students got either an "A" or "B". Let a be the number of students getting "A" and b be the number of students getting "B". Therefore, a and b are unknown parameters with $a + b = h$. Our goal is to use expectation maximization to obtain a maximum likelihood estimate of μ.

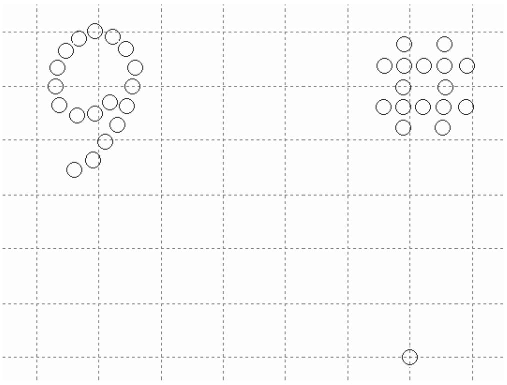
1. Use Multinoulli distribution to compute the log-likelihood function $l(\mu, a, b)$.
2. Expectation step: Given $\hat{\mu}^{(m)}$, compute the expected values $\hat{a}^{(m)}$ and $\hat{b}^{(m)}$ of a and b respectively.
3. Maximization step: Plug $\hat{a}^{(m)}$ and $\hat{b}^{(m)}$ into the log-likelihood function $l(\mu, a, b)$ and calculate for the maximum likelihood estimate $\hat{\mu}^{(m+1)}$ of μ, as a function of $\hat{\mu}^{(m)}$.
4. Iterating between the E-step and M-step will always converge to a local optimum of μ (which may or may not also be a global optimum)? Explain why in short.

Problem 4 (Spectral Clustering)

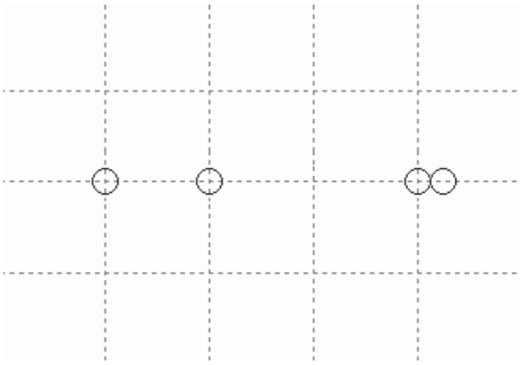
1. We consider the 2-clustering problem, in which we have N data points $x_{1:N}$ to be grouped in two clusters, denoted by A and B. Given the N by N affinity matrix W (**Remember that in class we define the affinity matrix in the way that the diagonal entries are zero for undirected graphs**), consider the following two problems:
   - Min-cut: minimize $\sum\limits_{i\in A}\sum\limits_{j\in B} W_{ij}$;
   - Normalized cut: minimize $\frac{\sum_{i\in A}\sum_{j\in B} W_{ij}}{\sum_{i\in A}\sum_{j=1}^{N} W_{ij}} + \frac{\sum_{i\in A}\sum j\in B W_{ij}}{\sum_{i=1}^{N}\sum_{j\in B} W_{ij}}$.

   a) The data points are shown in Figure (a) above. The grid unit is 1. Let $W_{ij} = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}$, give the clustering results of min-cut and normalized cut respectively (Please draw a rough sketch and give the separation boundary in the answer book).

   b) The data points are shown in Figure (b) above. The grid unit is 1. Let $W_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}}$, describe the clustering results of min-cut algorithm for $\sigma^2 = 50$ and $\sigma^2 = 0.5$ respectively (Please draw a rough sketch and give the separation boundaries for each case of σ^2 in the answer book).



(a)



(b)

2. Now back to the setting of the 2-clustering problem shown in Figure (a). The grid unit is 1.

   a) If we use Euclidean distance to construct the affinity matrix W as follows:

   $$W_{ij} = \begin{cases} 1, & \text{if } \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \leqslant \sigma^2; \\ 0, & \text{otherwise.} \end{cases}$$

   What σ^2 value would you choose? Briefly explain.

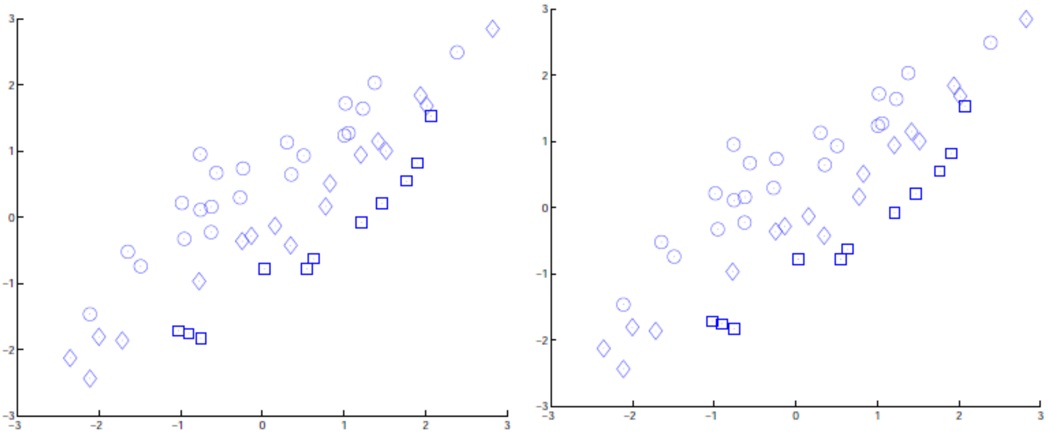
   b) The next step is to compute the first $k = 2$ dominant eigenvectors of the affinity matrix W. For the value of σ^2 you chose in the previous question, can you compute analytically the eigenvalues corresponding to the first two eigenvectors? If yes, compute and report the eigenvalues. If not, briefly explain.

✏ **Exercise 5** (Dimensionality Reduction)

1. (PCA vs. LDA) Plot the directions of the first PCA (plot (a)) and LDA (plot (b)) components in the following figures respectively.
2. (PCA and SVD) Given 6 data points in 5D space, $(1,1,1,0,0)$, $(-3,-3,-3,0,0)$, $(2,2,2,0,0)$, $(0,0,0,-1,-1)$, $(0,0,0,2,2)$, $(0,0,0,-1,-1)$. We can represent these data points by a 6×5 matrix $\mathbf{X}$, where each row corresponds to a data point:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ -3 & -3 & -3 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & -1 & -1 \end{pmatrix}$$

   a) What is the sample mean of the data set?



(c) First PCA component

(d) First LDA component

b) What is the SVD of the data matrix $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ satisfy $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_2$? Note that the SVD for this matrix must take the following form, where a, b, c, d, σ_1, σ_2 are the parameters you need to decide.

$$\mathbf{X} = \begin{pmatrix} a & 0 \\ -3a & 0 \\ 2a & 0 \\ 0 & b \\ 0 & -2b \\ 0 & b \end{pmatrix} \times \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \times \begin{pmatrix} c & c & c & 0 & 0 \\ 0 & 0 & 0 & d & d \end{pmatrix}$$

c) What is first principle component for the original data points?

d) If we want to project the original data points $\{\mathbf{x}_i\}_{i=1}^6$ into 1D space by principle component you choose, what is the sample variance of the projected data $\{\hat{\mathbf{x}}_i\}_{i=1}^6$?

e) For the projected data in d), now if we represent them in the original 5-d space, what is the reconstruction error $\frac{1}{6}\sum_{i=1}^6 \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2$?

✏ **Exercise 6**(PCA as factor analysis and SVD)

PCA of a set of data in $\mathbb{R}^p$ provide a sequence of best linear approximations to those data, of all ranks $q \leqslant p$. Denote the observations by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, and consider the rank-$q$ linear model for representing them

$$f(\alpha) = \mu + \mathbf{V}_q\alpha$$

where μ is a location vector in $\mathbb{R}^p$, V_q is a $p \times q$ matrix with q **orthogonal unit vectors** as columns, and α is a q vector of parameters. If we can find such a model, then we can reconstruct each $\mathbf{x}_i$ by a low dimensional coordinate vector α_i through

$$\mathbf{x}_i = f(\alpha_i) + \epsilon_i = \mu + \mathbf{V}_q\alpha_i + \epsilon_i \tag{1}$$

where $\epsilon_i \in \mathbb{R}^p$ are noise terms. Then PCA amounts to minimizing this reconstruction error by least square method

$$\min_{\mu,\{\alpha_i\},\mathbf{V}_q} \sum_{i=1}^{N} \|\mathbf{x}_i - \mu - \mathbf{V}_q \alpha_i\|^2$$

1. Assume $\mathbf{V}_q$ is known and treat μ and α_i as unknowns. Show that the least square problem
$$\min_{\mu,\{\alpha_i\}} \sum_{i=1}^{N} \|\mathbf{x}_i - \mu - \mathbf{V}_q \alpha_i\|^2$$
is minimized by
$$\hat{\mu} = \bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i, \tag{2}$$
$$\hat{\alpha}_i = \mathbf{V}_q^T(\mathbf{x}_i - \bar{\mathbf{x}}). \tag{3}$$
Also show that the solution for $\hat{\mu}$ is not unique. Give a family of solutions for $\hat{\mu}$.

2. For the standard solution (2), we are left with solving
$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} \|(\mathbf{x}_i - \bar{\mathbf{x}}) - \mathbf{V}_q\mathbf{V}_q^T(\mathbf{x}_i - \bar{\mathbf{x}})\|^2 = \min_{\mathbf{V}_q} \mathrm{Tr}\Big(\tilde{\mathbf{X}}(\mathbf{I}_p - \mathbf{V}_q\mathbf{V}_q^T)\tilde{\mathbf{X}}^T\Big). \tag{4}$$
Here we introduce the centered sample matrix
$$\tilde{\mathbf{X}} = (\mathbf{I}_N - \frac{1}{N}\mathbf{J}_N)\mathbf{X} = \begin{pmatrix} (\mathbf{x}_1 - \bar{\mathbf{x}})^T \\ \vdots \\ (\mathbf{x}_N - \bar{\mathbf{x}})^T \end{pmatrix} \in \mathbb{R}^{N\times p}$$
where $\mathbf{I}_N$ is $N\times N$ identity matrix and $\mathbf{J}_N$ is a matrix whose entries are all 1's. Recall the singular value decomposition (SVD) in linear algebra: $\tilde{\mathbf{X}} = \mathbf{U}\mathbf{D}\mathbf{V}^T$. Here $\mathbf{U}$ is an $N \times p$ orthogonal matrix ($\mathbf{U}^T\mathbf{U} = \mathbf{I}_p$) whose columns $\mathbf{u}_j$ are called the left singular vectors; $\mathbf{V}$ is a $p \times p$ orthogonal matrix ($\mathbf{V}^T\mathbf{V} = \mathbf{I}_p$) with columns $\mathbf{v}_j$ called the right singular vectors, and $\mathbf{D}$ is a $p \times p$ diagonal matrix, with diagonal elements $d_1 \geqslant d_2 \geqslant \cdots \geqslant d_p \geqslant 0$ known as the singular values.

   Show that the solution $\mathbf{V}_q$ to problem (4) consists of the first q columns of $\mathbf{V}$. (Then the optimal $\hat{\alpha}_i$ are given by the i-th row with the first q columns of $\mathbf{UD}$.)

Remark: The model (1), in general, gives the factor analysis in multivariate statistics:

$$\mathbf{x} = \mu + \mathbf{V}_q\alpha + \epsilon$$

In traditional factor analysis, α_j with $j = 1, \ldots, q$ is assumed to be Gaussian and uncorrelated as well as ϵ_i with $i = 1, \ldots, p$. However, Independent Component Analysis (ICA) instead assumes α_j with $j = 1, \ldots, q$ is assumed to be non-Gaussian and independent. Because of the independence, ICA is particularly useful in separating mixed signals.