Matthew Ragoza
Dr. Milos Hauskrecht
ISSP 2170 - Machine Learning
January 21, 2022

Assignment 1

Problem 1. Matrix operations

- $\mathbf{u}^T\mathbf{u} = 26$
- $\mathbf{u}\mathbf{u}^T = \begin{bmatrix} 16 & 4 & 12 \\ 4 & 1 & 3 \\ 12 & 3 & 9 \end{bmatrix}$
- $\mathbf{v}\mathbf{u} = 71$
- $\mathbf{u} + 5 = \begin{bmatrix} 9 \\ 6 \\ 8 \end{bmatrix}$
- $\mathbf{A}^T = \begin{bmatrix} 1 & 3 \\ 2 & 4 \\ 5 & 6 \end{bmatrix}$
- $\mathbf{B}\mathbf{u} = \begin{bmatrix} 56 \\ 19 \\ 42 \end{bmatrix}$
- $\mathbf{B}^{-1} = \begin{bmatrix} 1 & -11/2 & 5/4 \\ 0 & -1/2 & 1/4 \\ -2/3 & 13/3 & -1 \end{bmatrix}$
- $\mathbf{B} + \mathbf{C} = \begin{bmatrix} 15 & 7 & 14 \\ 3 & -1 & 7 \\ 3 & 6 & 10 \end{bmatrix}$
- $\mathbf{B} - \mathbf{C} = \begin{bmatrix} -1 & -5 & 4 \\ 1 & 5 & -1 \\ 5 & 10 & 2 \end{bmatrix}$
- $\mathbf{A}\mathbf{B} = \begin{bmatrix} 31 & 45 & 45 \\ 53 & 59 & 75 \end{bmatrix}$
- $\mathbf{B}\mathbf{C} = \begin{bmatrix} 48 & 21 & 75 \\ 15 & 0 & 30 \\ 34 & -12 & 76 \end{bmatrix}$
- $\mathbf{B}\mathbf{A}$ is undefined due to incompatible dimensions. Matrix multiplication defines a non-commutative mapping of $f : \mathbb{R}^{m\times n} \times \mathbb{R}^{n\times p} \to \mathbb{R}^{m\times p}$, but we have the matrices $\mathbf{B} \in \mathbb{R}^{3\times 3}$ and $\mathbf{A} \in \mathbb{R}^{2\times 3}$.

Problem 2. Exploratory data analysis

(a) The ranges of each attribute in the data are shown below in Table 1.

	preg	plas	pres	skin	test	mass	pedi	age	class
min	0	0	0	0	0	0.0	0.078	21	0
max	17	199	122	99	846	67.1	2.420	81	1

Table 1. Minimum and maximum value of each attribute.

(b) The mean and standard deviation of each attribute are shown in Table 2.

(c) The mean and standard deviation for each covariate, grouped by `class`, are shown in Table 3. The attribute I would use to distinguish the classes would be `plas`, because there is a difference in the means ($\bar{x}_0 = 109.98$ vs. $\bar{x}_1 = 141.26$) of over one standard deviation ($s = 31.97$) between the two classes. This is a larger difference between

	preg	plas	pres	skin	test	mass	pedi	age	class
mean	3.85	120.89	69.11	20.54	79.80	31.99	0.47	33.24	0.35
std	3.37	31.97	19.36	15.95	115.24	7.88	0.33	11.76	0.48

Table 2. Mean and standard deviation of each attribute.

class		preg	plas	pres	skin	test	mass	pedi	age
0	mean	3.30	109.98	68.18	19.66	68.79	30.30	0.43	31.19
	std	3.02	26.14	18.06	14.89	98.87	7.69	0.30	11.67
1	mean	4.87	141.26	70.82	22.16	100.34	35.14	0.55	37.07
	std	3.74	31.94	21.49	17.68	138.69	7.26	0.37	10.97

Table 3. Mean and standard deviation of each attribute by class.

the classes than in any other attribute. This implies that the classes are most easily separated using this variable, assuming that it follows a normal distribution.

(d) The correlations of each covariate in the data set with the class variable are shown in Table 4. The attribute that is most highly correlated with `class` is `plas`, which has a correlation coefficient of $r = 0.47$. Converting this to a coefficient of determination ($r^2 = 0.22$) tells us that 22% of class variance is explained by variance in `plas`. Out of all the variables, when `plas` varies, it is most likely that the class changes as well. For this reason, it is the variable that is *most* helpful in predicting the class.

	preg	plas	pres	skin	test	mass	pedi	age
class	0.22	**0.47**	0.07	0.07	0.13	0.29	0.17	0.24

Table 4. Correlation coefficients between class and other attributes.

(e) The correlation matrix between the different covariates in the data set is shown in Table 5. The two attributes that have the largest correlation ($r = 0.54, r^2 = 0.29$) are `preg` and `age`. This indicates that 29% of the variance in `preg` is explained by variance in `age`, and vice versa.

	preg	plas	pres	skin	test	mass	pedi	age
preg	1.00	0.13	0.14	-0.08	-0.07	0.02	-0.03	**0.54**
plas	0.13	1.00	0.15	0.06	0.33	0.22	0.14	0.26
pres	0.14	0.15	1.00	0.21	0.09	0.28	0.04	0.24
skin	-0.08	0.06	0.21	1.00	0.44	0.39	0.18	-0.11
test	-0.07	0.33	0.09	0.44	1.00	0.20	0.19	-0.04
mass	0.02	0.22	0.28	0.39	0.20	1.00	0.14	0.04
pedi	-0.03	0.14	0.04	0.18	0.19	0.14	1.00	0.03
age	**0.54**	0.26	0.24	-0.11	-0.04	0.04	0.03	1.00

Table 5. Pairwise correlation coefficients between all covariate attributes.

(f) Histograms for each of the covariates are displayed in Figure 1. The distributions of `preg`, `test`, `pedi` and `age` are all severely right-skewed. The histogram of `plas` is somewhat right-skewed and the ones for `pres` and `skin` have disproportionately many values in the zero bin. Out of all the attributes, the distribution that most resembles a normal distribution is that of `mass`. It is unimodal and nearly symmetric, with few outliers.

(g) Histograms for each of the covariates, grouped by `class`, are displayed in Figure 2. For nearly every variable, there is a large degree of overlap between the two classes. The one notable exception is `plas`. For this attribute, the histograms for the two classes have different centers and there are certain bins that have minor overlap. Specifically, the class representing individuals with diabetes tends to have higher values of `plas` that are very unlikely for those who do not have diabetes. Therefore, this attribute would be most useful for distinguishing individuals who fall into each class.

(h) Scatter plots for each pair of data covariates are displayed in Figure 3. If two variables are independent and random, with no significant correlation, we would expect the scatter plot to appear as a circular blob or cloud of points with no skew in any direction. One interesting pattern in the scatter plots is between `plas` and `test`. As `plas` increases, there is a tendency for `test` to increase as well, at least towards the upper end of the distribution. This indicates a positive correlation. Another pattern is between `age` and `pres`, where there is a similar trend. As `age` increases, the tendency is for `pres` to increase, implying a positive correlation.

Problem 3. Data preprocessing

(a) Using the following indexing function f, we can convert the provided color vector $\mathbf{v}$ to a one-hot encoded matrix $\mathbf{X}$, where $\mathbf{X}_{ij} = \mathbb{1}(f(\mathbf{v}_i) = j)$.

$$\mathbf{v} = \begin{bmatrix} red \\ black \\ yellow \\ red \\ green \\ blue \\ blue \end{bmatrix} \qquad f(x) = \begin{cases} 1 & x = brown \\ 2 & x = blue \\ 3 & x = white \\ 4 & x = red \\ 5 & x = yellow \\ 6 & x = orange \\ 7 & x = green \\ 8 & x = black \end{cases} \qquad \mathbf{X} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(b) After normalizing the values, the first five rows for attribute 3 (`skin`) are:

$$\begin{bmatrix} 0.907 & 0.530 & -1.288 & 0.154 & 0.907 \end{bmatrix}^T$$

(c) After discretizing the values, the first five rows for attribute 3 (`skin`) are:

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

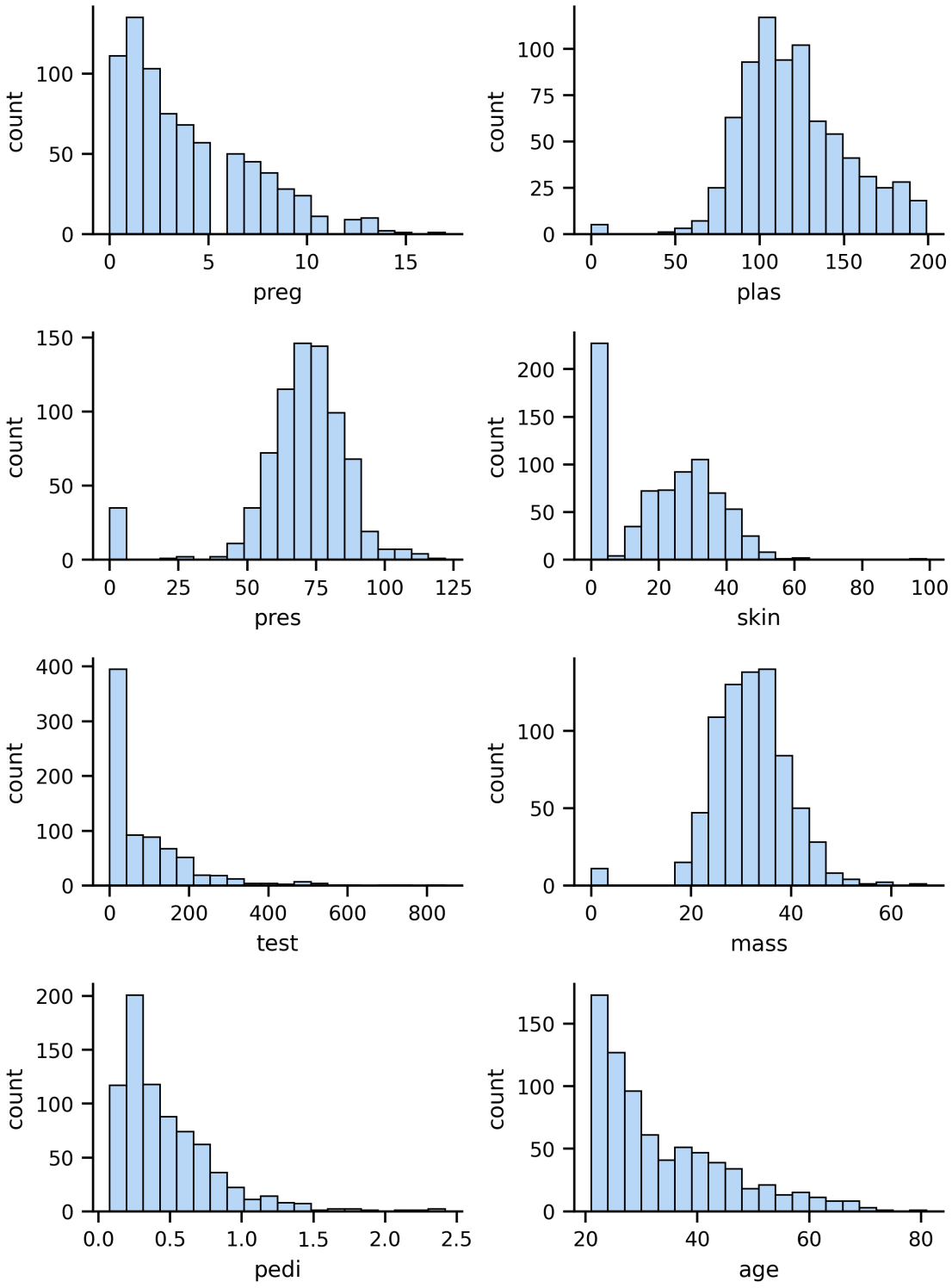

Figure 1. Histograms of each covariate in the data set.

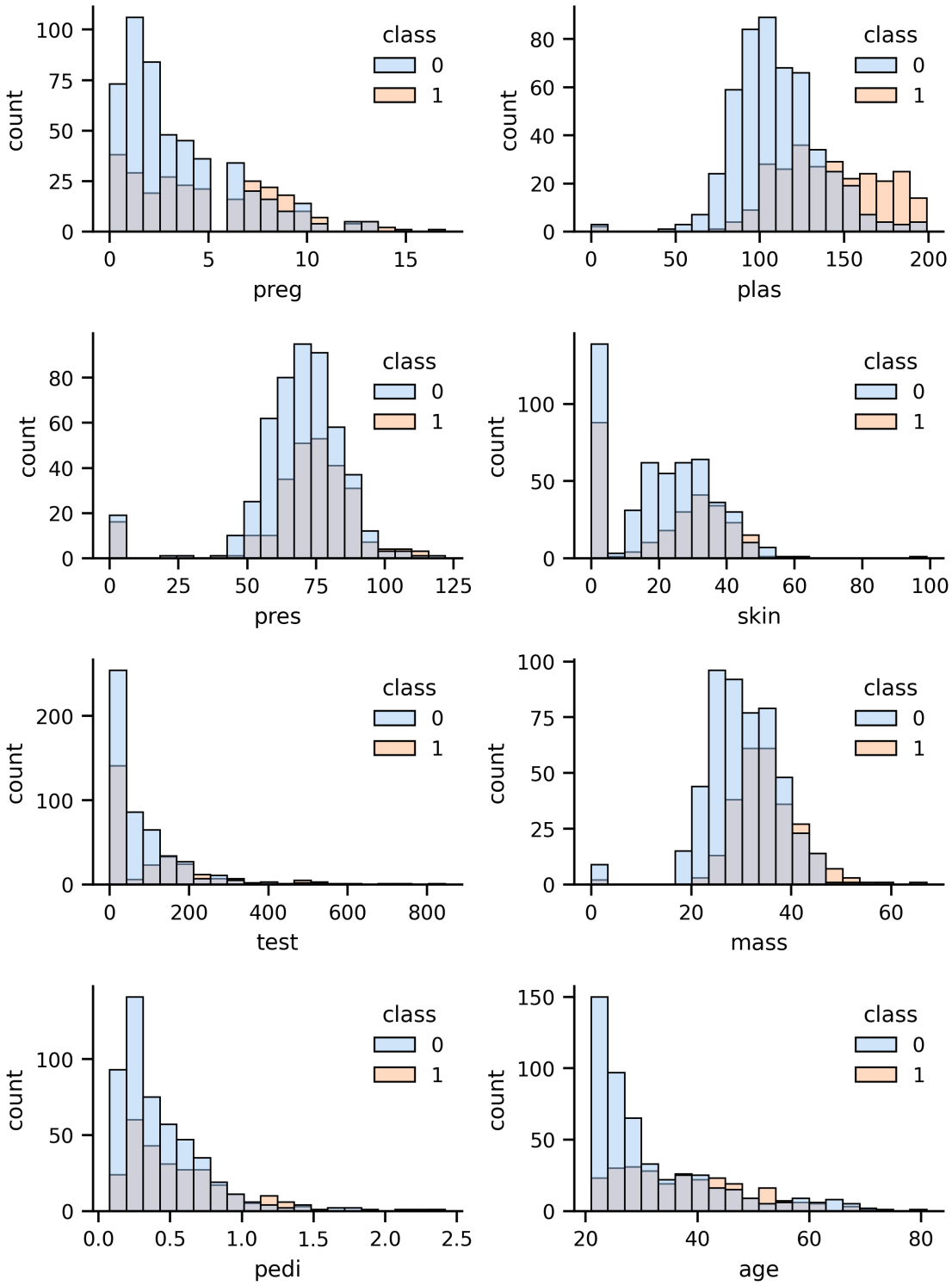

Figure 2. Histograms of each covariate in the data set, grouped by class.

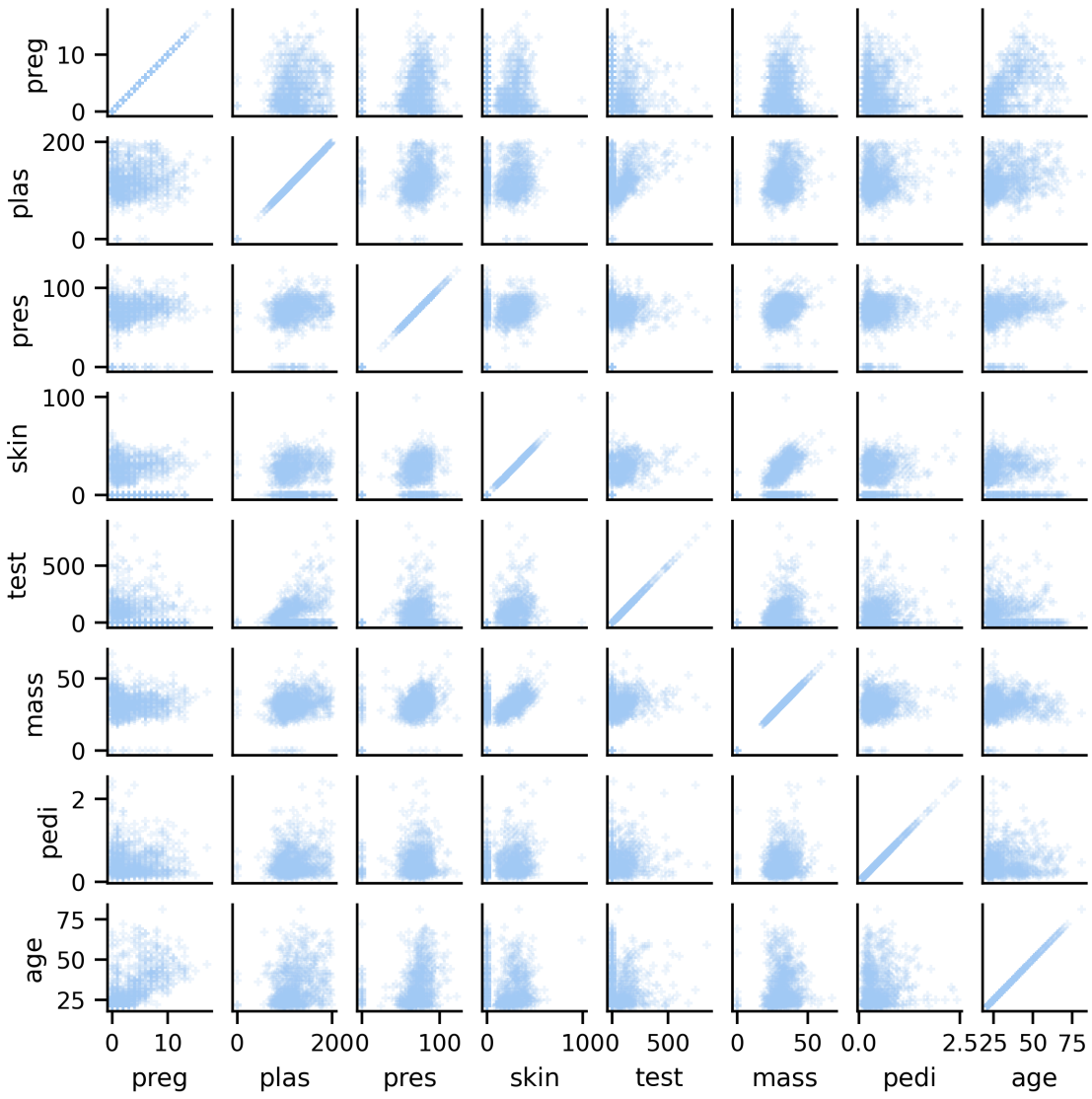

Figure 3. Pairwise scatter plots between each covariate in the data set.

Problem 4. Splitting into training and test sets

(a) There are 514 train and 254 test instances for `test_size=0.33` and `random_state=7`.
(b) There are 514 train and 254 test instances for `test_size=0.33` and `random_state=3`. The training and test sets contain different instances than in the previous split.
(c) There are 576 train and 192 test instances for `test_size=0.25` and `random_state=7`.