Machine Learning Department, Carnegie Mellon University

Linear Algebra Review

Jing Xiang

March 18, 2014

1 Properties of Matrices

Below are a few basic properties of matrices:

- Matrix Multiplication is associative: $(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC})$
- Matrix Multiplication is distributive: $\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}$
- Matrix Multiplication is NOT commutative in general, that is $\boldsymbol{AB} \neq \boldsymbol{BA}$. For example, if $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times q}$, the matrix product $\boldsymbol{BA}$ does not exist.

2 Transpose

The transpose of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, is written as $\boldsymbol{A}^\top \in \mathbb{R}^{n \times m}$ where the entries of the matrix are given by:

$$(\boldsymbol{A}^\top)_{ij} = \boldsymbol{A}_{ji} \tag{2.1}$$

Properties:

- Transpose of a scalar is a scalar $a^\top = a$
- $(\boldsymbol{A}^\top)^\top = \boldsymbol{A}$
- $(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top$
- $(\boldsymbol{A} + \boldsymbol{B})^\top = \boldsymbol{A}^\top + \boldsymbol{B}^\top$

3 Trace

The trace of a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is written as $\text{Tr}(\boldsymbol{A})$ and is just the sum of the diagonal elements:

$$\text{Tr}(\boldsymbol{A}) = \sum_{i=1}^{n} \boldsymbol{A}_{ii} \tag{3.1}$$

The trace of a product can be written as the sum of entry-wise products of elements.

$$\text{Tr}(\boldsymbol{A}^\top \boldsymbol{B}) = \text{Tr}(\boldsymbol{A}\boldsymbol{B}^\top) = \text{Tr}(\boldsymbol{B}^\top \boldsymbol{A}) = \text{Tr}(\boldsymbol{B}\boldsymbol{A}^\top) \tag{3.2}$$

$$= \sum_{i,j}^{n} \boldsymbol{A}_{i,j} \boldsymbol{B}_{i,j} \tag{3.3}$$

$$\tag{3.4}$$

Properties:

- Trace of a scalar is a scalar $\text{Tr}(a) = a$
- $\boldsymbol{A} \in \mathbb{R}^{n \times n}, \text{Tr}(\boldsymbol{A}) = \text{Tr}(\boldsymbol{A}^\top)$
- $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times n}, \text{Tr}(\boldsymbol{A} + \boldsymbol{B}) = \text{Tr}(\boldsymbol{A}) + \text{Tr}(\boldsymbol{B})$
- $\boldsymbol{A} \in \mathbb{R}^{n \times n}, c \in \mathbb{R}, \text{Tr}(c\boldsymbol{A}) = c\,\text{Tr}(\boldsymbol{A})$
- $\boldsymbol{A}, \boldsymbol{B}$ such that $\boldsymbol{AB}$ is square, $\text{Tr}(\boldsymbol{AB}) = \text{Tr}(\boldsymbol{BA})$
- $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ such that $\boldsymbol{ABC}$ is square, $\text{Tr}(\boldsymbol{ABC}) = \text{Tr}(\boldsymbol{BCA}) = \text{Tr}(\boldsymbol{CAB})$, this is called **trace rotation**.

4 Vector Norms

$\boldsymbol{A}$ norm of a vector $\|\mathbf{x}\|$ is a measure of it's "length" or "magnitude". The most common is the Euclidean or ℓ_2 norm.

1. ℓ_2 norm : $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$

   For example, this is used in ridge regression: $\|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_2^2$

2. ℓ_1 norm : $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$

   For example, this is used in ℓ_1 penalized regression: $\|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1$

3. ℓ_∞ norm : $\|\mathbf{x}\|_\infty = \max_i |x_i|$

4. The above are all examples of the family of ℓ_p norms : $\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$

5 Rank

A set of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n \subset \mathbb{R}^m$ is said to be linearly independent if no vector can be represented as a linear combination of the remaining vectors. The rank of a matrix is size of the largest subset of columns of A that constitute a linearly independent set. This is often referred to as the number of linearly independent columns of $\boldsymbol{A}$. Note the amazing fact that $\text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{A}^\top)$. This means that column rank = row rank.
For $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ $\text{rank}(\boldsymbol{A}) \leq \min(m, n)$. If $\text{rank}(\boldsymbol{A}) = \min(m, n)$, then $\boldsymbol{A}$ is full rank.

6 Inverse

The inverse of a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is written as $\boldsymbol{A}^{-1}$ and is defined such that:

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = I$$

If $\boldsymbol{A}^{-1}$ exists, the matrix is said to be **nonsingular**, otherwise it is **singular**. For a square matrix to be invertible, it must be full rank. Non-square matrices are not invertible.
Properties:

- $(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$
- $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$
- $(\boldsymbol{A}^{-1})^\top = (\boldsymbol{A}^\top)^{-1}$

Sherman-Morrison-Woodbury Matrix Inversion Lemma

$$(\boldsymbol{A} + \boldsymbol{X}\boldsymbol{B}\boldsymbol{X}^\top)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{X}(\boldsymbol{B}^{-1} + \boldsymbol{X}^\top\boldsymbol{A}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{A}^{-1}$$

This comes up and can often make a hard inverse into an easy inverse. $\boldsymbol{A}$ and $\boldsymbol{B}$ are square and invertible but they don't need to be the same dimension.

7 Orthogonal Matrices

- Two vectors are orthogonal if $\mathbf{u}^\top\mathbf{v} = 0$. A vector is normalized if $\|\mathbf{x}\| = 1$.
- A square matrix is orthogonal if all its columns are orthogonal to each other and are normalized (columns are orthonormal).
- If $\boldsymbol{U}$ is an orthogonal matrix $\boldsymbol{U}^\top = \boldsymbol{U}^{-1}$, then $\boldsymbol{U}^\top\boldsymbol{U} = I = \boldsymbol{U}\boldsymbol{U}^\top$.
- Note if $\boldsymbol{U}$ is not square, but the columns are orthonormal, then $\boldsymbol{U}^\top\boldsymbol{U} = I$ but $\boldsymbol{U}\boldsymbol{U}^\top \neq I$. Orthogonal usually refers to the first case.

8 Matrix Calculus

Gradient

Given $f : \mathbb{R}^{m \times n} \to \mathbb{R}$ is a function that takes as input a matrix and returns a real values. Then the gradient of f with respect to $\boldsymbol{A}$ is the matrix of partial derivatives, that is the $m \times n$ matrix defined below.

$$(\bigtriangledown_{\boldsymbol{A}} f(\boldsymbol{A}))_{ij} = \frac{\partial f(\boldsymbol{A})}{\partial A_{ij}}$$

Note that the size of $\bigtriangledown_{\boldsymbol{A}} f(\boldsymbol{A}))_{ij}$ is the same as the size of $\boldsymbol{A}$.
The gradient of a vector $\mathbf{x} \in \mathbb{R}^n$ is the following:

$$(\bigtriangledown_{\mathbf{x}} f(\mathbf{x}))_i = \frac{\partial f(\mathbf{x})}{\partial x_i}$$

The gradient of a function is only defined if that function is real-valued, that is it returns a real scalar value.

Hessian

Given $f : \mathbb{R}^n \to \mathbb{R}$ is a function that takes a vector and returns a real number. Then the Hessian of f with respect to $\mathbf{x}$ is a $n \times n$ matrix of partial derivatives as defined below.

$$(\bigtriangledown^2_{\mathbf{x}} f(\mathbf{x}))_{ij} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}$$

Just like the gradient, the Hessian is only defined when the function is real-valued. For the purposes of this class, we will only be taking the Hessian of a vector.

Common forms of Derivatives

$$\frac{\partial(\mathbf{a}^\top \mathbf{x})}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}^\top \mathbf{a})}{\partial \mathbf{x}} = \mathbf{a}$$

$$\frac{\partial(\mathbf{x}^\top \boldsymbol{A} \mathbf{x})}{\partial \mathbf{x}} = (\boldsymbol{A} + \boldsymbol{A}^\top)\mathbf{x}$$

$$\frac{\partial(\mathbf{a}^\top \boldsymbol{X} \mathbf{b})}{\partial \boldsymbol{X}} = \mathbf{a}\mathbf{b}^\top$$

$$\frac{\partial(\mathbf{a}^\top \boldsymbol{X}^\top \mathbf{b})}{\partial \boldsymbol{X}} = \mathbf{b}\mathbf{a}^\top$$

$$\frac{\partial(\mathbf{a}^\top \boldsymbol{X} \mathbf{a})}{\partial \boldsymbol{X}} = \frac{\partial(\mathbf{a}^\top \boldsymbol{X}^\top \mathbf{a})}{\partial \boldsymbol{X}} = \mathbf{a}\mathbf{a}^\top$$

$$
\begin{aligned}
\frac{\partial(\mathbf{x}^\top \boldsymbol{A})}{\partial \mathbf{x}} &= \boldsymbol{A} \\
\frac{\partial(\mathbf{x}^\top)}{\partial \mathbf{x}} &= \boldsymbol{I} \\
\frac{\partial(\boldsymbol{A}\mathbf{x})}{\partial z} &= \boldsymbol{A}\frac{\partial \mathbf{x}}{\partial z} \\
\frac{\partial(\boldsymbol{X}\boldsymbol{Y})}{\partial z} &= \boldsymbol{X}\frac{\partial \boldsymbol{Y}}{\partial z} + \frac{\partial \boldsymbol{X}}{\partial z}\boldsymbol{Y} \\
\frac{\partial(\boldsymbol{X}^{-1})}{\partial z} &= -\boldsymbol{X}^{-1}\frac{\partial \boldsymbol{X}}{\partial z}\boldsymbol{X}^{-1} \\
\frac{\partial \ln|\boldsymbol{X}|}{\partial \boldsymbol{X}} &= (\boldsymbol{X}^{-1})^\top = (\boldsymbol{X}^\top)^{-1}
\end{aligned}
$$

9 Linear Regression

To begin, the likelihood can be derived from a multivariate normal distribution. The likelihood for linear regression is given by:

$$
\begin{aligned}
P(\mathcal{D}|\boldsymbol{\beta}, \sigma^2) = P(\mathbf{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^{n} \mathcal{N}(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})\right)
\end{aligned}
$$

By taking the log and throwing away constants, we get the negative log-likelihood below.

$$
-\log P(\mathcal{D}|\boldsymbol{\beta}, \sigma^2) = \frac{n}{2}\log(\sigma^2) + \frac{1}{2\sigma^2}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})
$$

We can now define the residual sum of squares or least squares.

$$
\|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\| = (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})
$$

Maximizing the likelihood is equivalent to minimizing the negative log likelihood and also equivalent to minimizing the residual sum of squares. You will also hear this being called finding the least squares solution. We can rewrite the expression as follows.

$$
\begin{aligned}
\|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\| &= (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= \mathbf{y}^\top\mathbf{y} - 2(\boldsymbol{X}^\top\mathbf{y})^\top\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}
\end{aligned}
$$

To find the minimum, we first have to take the derivative. Note, we need two matrix derivative identities $\frac{\partial \mathbf{x}^\top \boldsymbol{A}\mathbf{x}}{\partial \mathbf{x}} = (\boldsymbol{A} + \boldsymbol{A}^\top)\mathbf{x}$ and $\frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$. Also, note that $\boldsymbol{X}^\top \boldsymbol{X}$ is symmetric.

$$\begin{aligned}
&\frac{\partial(\mathbf{y}^\top \mathbf{y} - 2(\boldsymbol{X}^\top \mathbf{y})^\top \boldsymbol{\beta} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&= -2(\boldsymbol{X}^\top \mathbf{y}) + (\boldsymbol{X}^\top \boldsymbol{X} + (\boldsymbol{X}^\top \boldsymbol{X})^\top)\boldsymbol{\beta} \\
&= -2(\boldsymbol{X}^\top \mathbf{y}) + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}
\end{aligned}$$

After setting the derivation equal to zero and solving for $\boldsymbol{\beta}$, we get the following.

$$\begin{aligned}
0 &= -2(\boldsymbol{X}^\top \mathbf{y}) + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \\
\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} &= \boldsymbol{X}^\top \mathbf{y} \\
\boldsymbol{\beta} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathbf{y}
\end{aligned}$$

These are called the normal equations. To solve this in Octave/Matlab, you can implement the equations explicitly using the inverse. However, doing `beta = X \ y;` is a more stable way of solving the normal equations. It does a QR decomposition.

You can check that this solution is the global minimum and not just a stationary point. To do this, you need to evaluate the Hessian, or the second derivative. You should find that the result is a positive definite matrix. And since the Hessian is positive definite, the function is convex and thus the only stationary point is also the global minimum.

10 Ridge Regression

Now, we're going to derive ridge regression in a similar way. Recall that for linear regression, we found the MLE from forming the likelihood $P(\mathbf{y}|\boldsymbol{\beta})$. Here, we can derive the MAP estimate from the posterior which is constructed from the likelihood and the prior. Let $\boldsymbol{\beta} \sim \mathcal{N}(0, \frac{1}{\lambda}\boldsymbol{I}_p)$ be a prior on the parameter vector $\boldsymbol{\beta}$ where $\boldsymbol{I}_p$ is an identity matrix of size p. The form of the posterior is given below.

$$\begin{aligned}
P(\boldsymbol{\beta}|\mathbf{y}) &\propto P(\mathbf{y}|\boldsymbol{\beta})P(\boldsymbol{\beta}) \\
&\propto \mathcal{N}(\mathbf{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)\mathcal{N}(0, \frac{1}{\lambda}\boldsymbol{I}_p)
\end{aligned}$$

Given that $\sigma^2 = 1$, we first want to derive the posterior for $\boldsymbol{\beta}$.

$$
\begin{aligned}
P(\boldsymbol{\beta}|\mathbf{y}) &\propto P(\mathbf{y}|\boldsymbol{\beta})P(\boldsymbol{\beta}) \\
&\propto \mathcal{N}(\mathbf{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)\mathcal{N}(0, \frac{1}{\lambda}\boldsymbol{I}_p) \\
&\propto (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})\right) \cdot (2\pi)^{-\frac{p}{2}} \left|\frac{1}{\lambda}\boldsymbol{I}_p\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\beta}^\top(\frac{1}{\lambda}\boldsymbol{I}_p)^{-1}\boldsymbol{\beta}\right) \\
&\propto (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})\right) \cdot (2\pi)^{-\frac{p}{2}} (\frac{1}{\lambda})^{-\frac{p}{2}} \exp\left(-\frac{\lambda}{2}\boldsymbol{\beta}^\top\boldsymbol{I}_p\boldsymbol{\beta}\right) \\
&\propto (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})\right) \cdot (2\pi\lambda^{-1})^{-\frac{p}{2}} \exp\left(-\frac{\lambda}{2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right) \\
&\propto (2\pi\sigma^2)^{-\frac{n}{2}}\ (2\pi\lambda^{-1})^{-\frac{p}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) - \frac{\lambda}{2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right)
\end{aligned}
$$

Taking the negative log and dropping constants, we get:

$$
\begin{aligned}
&\propto \frac{1}{2\sigma^2}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{\lambda}{2}\boldsymbol{\beta}^\top\boldsymbol{\beta} \\
&\propto (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} \qquad \text{Setting } \sigma^2 \text{ to 1 and dropping more constants.}
\end{aligned}
$$

Now, since we wanted to maximize the posterior, we now need to minimize the negative log of the posterior. Note that minimizing the above expression is exactly the same as finding the ridge solution by minimizing the sum of squares plus the l_2 penalty (Eq. 10.1). These two expressions are equivalent, and thus minimizing them will yield identical solutions.

$$
\|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\| + \lambda\|\boldsymbol{\beta}\|_2^2 = (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} \tag{10.1}
$$

Let's expand out and write the loss function in matrix form.

$$
\begin{aligned}
&(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} \\
&= \mathbf{y}^\top\mathbf{y} - 2(\boldsymbol{X}^\top\mathbf{y})^\top\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}
\end{aligned}
$$

To find the value of $\boldsymbol{\beta}$ that minimizes the loss function, we first have to take the derivative. Note, we need two matrix derivative identities $\frac{\partial \mathbf{x}^\top\boldsymbol{A}\mathbf{x}}{\partial \mathbf{x}} = (\boldsymbol{A} + \boldsymbol{A}^\top)\mathbf{x}$ and $\frac{\partial \mathbf{a}^\top\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$. Also, note that $\boldsymbol{X}^\top\boldsymbol{X}$ is symmetric.

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\beta}}(\mathbf{y}^\top \mathbf{y} - 2(\boldsymbol{X}^\top \mathbf{y})^\top \boldsymbol{\beta} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}) \\
&= -2(\boldsymbol{X}^\top \mathbf{y}) + (\boldsymbol{X}^\top \boldsymbol{X} + (\boldsymbol{X}^\top \boldsymbol{X})^\top)\boldsymbol{\beta} + 2\lambda \boldsymbol{\beta} \\
&= -2\boldsymbol{X}^\top \mathbf{y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} + 2\lambda \boldsymbol{\beta}
\end{aligned}
$$

After setting the derivation equal to zero and solving for $\boldsymbol{\beta}$, we get the following.

$$
\begin{aligned}
0 &= -2\boldsymbol{X}^\top \mathbf{y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} + 2\lambda \boldsymbol{\beta} \\
\boldsymbol{X}^\top \mathbf{y} &= \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} + \lambda \boldsymbol{\beta} \\
\boldsymbol{X}^\top \mathbf{y} &= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})\boldsymbol{\beta} \\
\boldsymbol{\beta} &= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \mathbf{y}
\end{aligned}
$$

Just like linear regression, you can implement the equations explicitly in Matlab/Octave. In practice, you might have trouble calculating the inverse directly if the matrix is huge and λ is small. We can also derive a numerically stable way of computing $\boldsymbol{\beta}$ using the backslash operator. Define $\tilde{\boldsymbol{X}}$ and $\tilde{\mathbf{y}}$ such that $\boldsymbol{\beta}$ can be written as:

$$
\boldsymbol{\beta} = (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}} \tilde{\mathbf{y}} \tag{10.2}
$$

Then, you can use the backslash operator as shown below.

```
Xtil = [X; sqrt(lambda)*eye(p)];
ytil=[y; zeros(p,1)];
beta = Xtil\ytil;
```

11 Quadratic Forms

For a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ and a vector $\mathbf{x} \in \mathbb{R}^n$, the scalar value $\mathbf{x}^\top \boldsymbol{A}\mathbf{x}$ is referred to as quadratic form. We can write it explicitly as follows:

$$
\mathbf{x}^\top \boldsymbol{A}\mathbf{x} = \sum_{i=1}^{n} x_i (\boldsymbol{A}\mathbf{x})_i = \sum_{i=1}^{n} x_i \left(\sum_{j=1}^{n} A_{ij} x_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j
$$

11.1 Definitions

Positive Definite (PD)
notation: $\boldsymbol{A} > 0$ or $\boldsymbol{A} \succ 0$ and the set of all positive definite matrices $\mathbb{S}^n_{++}$.
A symmetric matrix $\boldsymbol{A} \in \mathbb{S}^n$ is positive definite if for all non-zero vectors $\mathbf{x} \in \mathbb{R}, \mathbf{x}^\top \boldsymbol{A}\mathbf{x} > 0$.

Positive Semidefinite (PSD) notation: $\boldsymbol{A} \geq 0$ or $\boldsymbol{A} \succeq 0$ and the set of all positive semidefinite matrices $\mathbb{S}^n_{+}$.
A symmetric matrix $\boldsymbol{A} \in \mathbb{S}^n$ is positive semidefinite if for all non-zero vectors $\mathbf{x} \in \mathbb{R}, \mathbf{x}^\top \boldsymbol{A}\mathbf{x} \geq 0$.

Negative Definite (ND) notation: $\boldsymbol{A} < 0$ or $\boldsymbol{A} \prec 0$.
Similarly, a symmetric matrix $\boldsymbol{A} \in \mathbb{S}^n$ is negative definite if for all non-zero vectors $\mathbf{x} \in \mathbb{R}, \mathbf{x}^\top \boldsymbol{A}\mathbf{x} < 0$.

Negative Semidefinite (NSD) notation: $\boldsymbol{A} \leq 0$ or $\boldsymbol{A} \preceq 0$.
Similarly, a symmetric matrix $\boldsymbol{A} \in \mathbb{S}^n$ is negative semidefinite if for all non-zero vectors $\mathbf{x} \in \mathbb{R}, \mathbf{x}^\top \boldsymbol{A}\mathbf{x} \leq 0$.

Indefinite Lastly, a symmetric matrix $\boldsymbol{A} \in \mathbb{S}^n$ is indefinite if it is neither positive semidefinite nor negative semidefinite, that is if there exists $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}$ such that $\mathbf{x}_1^\top \boldsymbol{A}\mathbf{x}_1 > 0$ and $\mathbf{x}_2^\top \boldsymbol{A}\mathbf{x}_2 < 0$.

If $\boldsymbol{A}$ is positive definite, then $-\boldsymbol{A}$ is negative definite and vice versa. The same can be same about positive semidefinite and negative semidefinite. Also, positive definite and negative definite matrices are always full rank and invertible.

12 Eigenvalues and Eigenvectors

Given a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}, \lambda \in \mathbb{C}$ is an **eigenvalue** and $\mathbf{x} \in \mathbb{C}$ (complex set of numbers) the corresponding eigenvector if

$$\boldsymbol{A}\mathbf{x} = \lambda \mathbf{x}, \ \mathbf{x} \neq 0$$

This condition can be rewritten as:

$$(\boldsymbol{A} - \lambda \boldsymbol{I})\mathbf{x} = 0$$

where $\boldsymbol{I}$ is the identity matrix. Now for a non-zero vector to satisfy this equation, then $(\boldsymbol{A} - \lambda \boldsymbol{I})$ must not be invertible, which means that it is singular and the determinant is zero.

You can use the definition of the determinant to expand this expression into a polynomial in λ and then find the roots (real or complex) of the polynomial to find the n eigenvalues $\lambda_1, \ldots, \lambda_n$. Once you have the eigenvalues λ_i, you can find the corresponding eigenvector by solving the system of equations $(\lambda_i \boldsymbol{I} - \boldsymbol{A})\mathbf{x} = 0$.

12.1 Properties

- The trace of a matrix $\boldsymbol{A}$ is equal to the sum of its eigenvalues:

$$\text{Tr}(\boldsymbol{A}) = \sum_{i=1}^{n} \lambda_i$$

- The determinant of $\boldsymbol{A}$ is equal to the product of its eigenvalues

$$|\boldsymbol{A}| = \prod_{i=1}^{n} \lambda_i$$

- The rank of $\boldsymbol{A}$ is equal to the number of non-zero eigenvalues of $\boldsymbol{A}$
- The eigenvalues of a diagonal matrix $\boldsymbol{D} = \text{diag}(d_1, \ldots, d_n)$ are just the diagonal entries $d_1, \ldots, d_n$

12.2 Diagonalization

A square matrix $\boldsymbol{A}$ is said to be diagonalizable if it is similar to a diagonal matrix. A diagonal matrix $\boldsymbol{A}$ has the property that there exists an invertible matrix $\boldsymbol{X}$ and a diagonal matrix Λ such that $\boldsymbol{A} = \boldsymbol{X}\Lambda\boldsymbol{X}^{-1}$.

We can write all the eigenvector equations simultaneously as $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{X}\Lambda$ where the columns of $\boldsymbol{X} \in \mathbb{R}^{n \times n}$ are the eigenvectors of A and Λ is a diagonal matrix whose entries are the eigenvalues of $\boldsymbol{A}$. If the eigenvectors of $\boldsymbol{A}$ are linearly independent, then the matrix $\boldsymbol{X}$ will be invertible, so $\boldsymbol{A} = \boldsymbol{X}\Lambda\boldsymbol{X}^{-1}$. This is known as the **eigenvalue decomposition** of the matrix.

Why is this useful? Because powers of diagonal matrices are easy to compute. Try computing $\boldsymbol{A}^3$. Also, remember this form $\boldsymbol{A} = \boldsymbol{X}\Lambda\boldsymbol{X}^{-1} = \boldsymbol{X}\Lambda\boldsymbol{X}^{\top} = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i^{\top}$. We will see this later when we cover SVMs with kernels: $\sum_{i=1}^{n} \lambda_i \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^{\top}$.

12.3 Properties of Eigenvalues/Eigenvectors for Symmetric Matrices

- For a symmetric matrix $\boldsymbol{A} \in \mathbb{S}^n$, all the eigenvalues are real.
- The eigenvectors of A are orthonormal so that means the matrix $\boldsymbol{X}$ is an orthogonal matrix (so we can denote the matrix of eigenvectors as $\boldsymbol{U}$).

We can then write

$$\boldsymbol{A} = \boldsymbol{X}\Lambda\boldsymbol{X}^{-1}$$
$$\boldsymbol{A} = \boldsymbol{U}\Lambda\boldsymbol{U}^{-1} \text{ The inverse of an orthogonal matrix is just the inverse.}$$
$$\boldsymbol{A} = \boldsymbol{U}\Lambda\boldsymbol{U}^{\top}$$

This means that

$$\begin{aligned}\mathbf{x}^{\top}\boldsymbol{A}\mathbf{x} &= \mathbf{x}^{\top}\boldsymbol{U}\Lambda\boldsymbol{U}^{\top}\mathbf{x} \\ &= \mathbf{y}^{\top}\Lambda\mathbf{y} \\ &= \sum_{i=1}^{n}\lambda_i y_i^2\end{aligned}$$

Since y_i^2 is always positive, the sign of this expression depends entirely on the $\lambda_i' s$. If all $\lambda_i > 0$, then the matrix is positive definite; if all $\lambda_i \geq 0$, then $\boldsymbol{A}$ is positive semidefinite. If $\lambda_i < 0$ and $\lambda_i \leq 0$, then the matrix is negative definite or negative semidefinite respectively. If $\boldsymbol{A}$ has both positive and negative eigenvalues, then it is indefinite.

13 Singular Value Decomposition

Any $n \times m$ matrix $\boldsymbol{A}$ can be written as

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}$$

where

$$\begin{aligned}\boldsymbol{U} &= \text{eigenvectors of } \boldsymbol{A}\boldsymbol{A}^{\top} \ (n \times n) \\ \boldsymbol{\Sigma} &= \sqrt{\text{diag}(\text{eig}(\boldsymbol{A}\boldsymbol{A}^{\top}))} \ (n \times m) \\ \boldsymbol{V} &= \text{eigenvectors of } \boldsymbol{A}^{\top}\boldsymbol{A} \ (m \times m)\end{aligned}$$

13.1 Properties

$$\boldsymbol{U}^{\top}\boldsymbol{U} = \boldsymbol{I}$$
$$\boldsymbol{U}\boldsymbol{U}^{\top} = \boldsymbol{I}$$
$$\boldsymbol{V}^{\top}\boldsymbol{V} = \boldsymbol{I}$$
$$\boldsymbol{V}\boldsymbol{V}^{\top} = \boldsymbol{I}$$

However, if you do the economy SVD, all the above properties are true except $\boldsymbol{U}\boldsymbol{U}^{\top} \neq 0$.

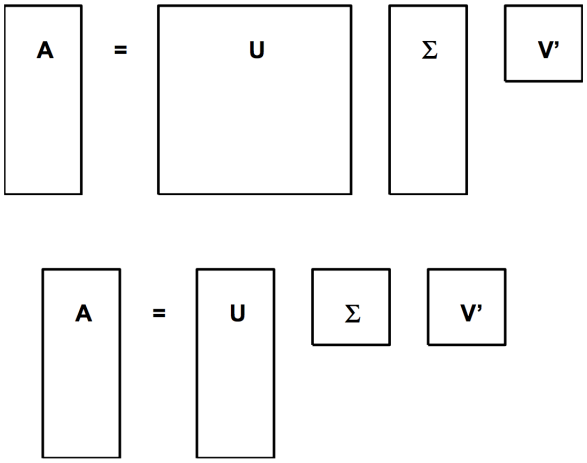

Figure 13.1: Taken from Matrix Cookbook.

13.2 Relation to Eigenvalue Decomposition

$$\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^\top \boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^\top$$
$$\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{V}\boldsymbol{\Sigma}^\top \boldsymbol{U}^\top = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^\top$$

The columns of $\boldsymbol{V}$ are the eigenvectors of $\boldsymbol{A}^\top\boldsymbol{A}$.
The columns of $\boldsymbol{U}$ are the eigenvectors of $\boldsymbol{A}\boldsymbol{A}^\top$.
The values of $\boldsymbol{\Sigma}, \sigma_i$ are the square roots of the eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$ or $\boldsymbol{A}\boldsymbol{A}^\top$, so $\sigma_i = \sqrt{\lambda_i}$

14 Principal Components Analysis

Often times when we have data in high-dimensional space, we can actually reduce the dimensions considerably while still capturing most of the variance of the data. This is called dimensionality reduction and one of the approaches is to use principal component analysis or PCA. PCA basically approximates some real $m \times n$ matrix $\boldsymbol{A}$ with he sum of some simple matrices that are rank one outer products.

The SVD of matrix A can be written:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$

where

$$\boldsymbol{A} = \boldsymbol{E}_1 + \boldsymbol{E}_2 + \cdots + \boldsymbol{E}_p,$$

where $p = \min(m, n)$. The component matrices $\boldsymbol{E}_i$ are rank one outer products:

$$\boldsymbol{E}_i = \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

The component matrices are orthogonal to each other so, the product is 0.

$$\boldsymbol{E}_j \boldsymbol{E}_k^\top = 0, \text{ where } j \neq k$$

The norm of each component matrix is the corresponding singular value.

$$\|\boldsymbol{E}\|_i = \sigma_i$$

So, the contribution that each component makes to reproducing A is determined by the size of the singular value. So, if you wanted to figure out how many components to include, you can plot the singular values and then cut it off where there is a significant drop in the value.

15 References

The following are my sources for this tutorial and you should check them out for further reading.

Zico Kolter's Linear Algebra Review and Reference
`http://cs229.stanford.edu/section/cs229-linalg.pdf`

The Matrix Cookbook
`http://orion.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`

Matlab's Eigenvalues and Singular Values
`http://www.mathworks.com/moler/eigs.pdf`

Course Notes from Harvard on Eigenvalues and Eigenvectors
`http://www.math.harvard.edu/archive/20_spring_05/handouts/ch05_notes.pdf`

Machine Learning: A Probabilistic Perspective by Kevin Murphy
`http://www.cs.ubc.ca/~murphyk/MLbook/index.html`