Matrix Methods in Machine Learning
Lecture Notes

Rebekah Dix

October 2, 2018

Contents

1 Elements of Machine Learning

1. Collect data
2. Preprocessing: changing data to simplify subsequent operations without losing relevant information.
3. Feature extraction: reduce raw data by extracting features or properties relevant to the model.
4. Generate training samples: a large collection of examples we can use to learn the model.
5. Loss function: To learn the model, we choose a loss function (i.e. a measure of how well a model fits the data)
6. Learn the model: Search over a collection of candidate models or model parameters to find one that minimizes the loss on training data.
7. Characterize generalization error (the error of our predictions on new data that was not used for training).

2 Linear Algebra Review

2.1 Products

Inner products:

$$\langle x, w \rangle = \sum_{j=1}^{p} w_j x_j = x^T w = w^T x \tag{1}$$

Thus this inner product is a weighted sum of the elements of x.

Matrix-vector multiplication:

$$Xw = \begin{bmatrix} \text{—} x_1^T \text{—} \\ \text{—} x_2^T \text{—} \\ \vdots \\ \text{—} x_n^T \text{—} \end{bmatrix} w = \begin{bmatrix} x_1^T w \\ x_2^T w \\ \vdots \\ x_n^T w \end{bmatrix} \tag{2}$$

Matrix-matrix multiplication:

Example 1. Let $X \in \mathbb{R}^{n \times p}$, n movies, p people. $T \in \mathbb{R}^{n \times r}$, and $W \in \mathbb{R}^{r \times p}$. We can think of T as the taste profiles of r representative customers and W as the weights on each representative profile (there will be one set of weights for each customer). Suppose we have two representative taste profiles (i.e. an action lover and a romance lover). Then w will be a 2-vector containing the weights of on the two representative taste profiles. Then

Tw is the expected preferences of a customer who weights the representative taste profiles of T with the weights given in w.

Now we can think about the full matrix product $X = TW$

$$X = TW \implies X_{ij} = \langle i\text{th row of T}, j\text{th column of W}\rangle \tag{3}$$

- The jth column of X is a weighted sum of the columns of T, where the jth column of W tells us the weights.
$$x_j = Tw_j \tag{4}$$
That is, the tastes (preferences) of the jth customer.
- The ith row of X is $x_i^T = t_i^T W$ where t_i^T is the ith row of T. This gives us how much each customer likes movie i.

Inner product representation:

$$TW = \begin{bmatrix} —t_1^T— \\ —t_2^T— \\ \vdots \\ —t_n^T— \end{bmatrix} \begin{bmatrix} | & | & & | \\ w_1 & w_2 & \dots & w_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} t_1^T w_1 & t_1^T w_2 & \dots & t_1^T w_p \\ t_2^T w_1 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ t_n^T w_1 & & & t_n^T w_p \end{bmatrix} \tag{5}$$

Outer Product Representation:

$$TW = \begin{bmatrix} | & | & & | \\ T_1 & T_2 & \dots & T_r \\ | & | & & | \end{bmatrix} \begin{bmatrix} —w_1^T— \\ —w_2^T— \\ \vdots \\ —w_r^T— \end{bmatrix} = \sum_{k=1}^{r} T_k w_k^T \tag{6}$$

(the sum of rank 1 matrices. TW has rank r if and only if the columns of T are rows of W are linearly independent). In this representation, we can think about T_k as the kth representative taste profile and w_k^T as the kth row of W, or the affinity of each customer with the kth representative profile.

2.2 Linear Independence

Definition 1. *(Linear Independence) Vectors $v_1, v_2, \dots, v_n \in \mathbb{R}^p$ are linearly independent vectors if and only if*

$$\sum_{j=1}^{n} \alpha_j v_j = 0 \iff \alpha_j = 0, j = 1, \dots, n \tag{7}$$

Definition 2. *(Matrix rank) The rank of a matrix is the maximum number of linearly independent columns. The rank of a matrix is less than the smallest dimension of the matrix.*

3 Linear Systems and Vector Norms

4 Least Squares

4.1 Vector Calculus Approach

4.1.1 Review of Vector Calculus

Let w be a p-vector and let f be a function of w that maps $\mathbb{R}^p$ to $\mathbb{R}$. Then the gradient of f with respect to w is

$$\nabla_w f(w) = \begin{pmatrix} \frac{\partial f(w)}{\partial w_1} \\ \vdots \\ \frac{\partial f(w)}{\partial w_p} \end{pmatrix} \tag{8}$$

Example 2. (Gradient of an Inner Product) Let $f(w) = \langle a, w \rangle = w^T a = \sum_{i=1}^n w_i a_i$. Then

$$\nabla_w w^T a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = a \tag{9}$$

Example 3. (Gradient of an Inner Product, Squared) Let $f(w) = \|w\|^2 = w^T w = w_1^2 + \cdots + w_p^2$. Then

$$\nabla_w w^T w = \begin{pmatrix} 2w_1 \\ 2w_2 \\ \vdots \\ 2w_p \end{pmatrix} = 2w \tag{10}$$

(This is a special case of the Quadratic Form discussed below, where $w^T Q w$, and $Q = I$)

Example 4. (Gradient of a Quadratic Form) Let $x \in \mathbb{R}^n$ and $f(x) = x^T Q x$, where Q is symmetric (if Q isn't symmetric we could replace Q with $\frac{1}{2}(Q + Q^T)$). Then

$$\begin{aligned} f(x) &= x^T Q x \\ &= \sum_{i=1}^n \sum_{j=1}^n x_i Q_{ij} x_j \end{aligned}$$

Therefore

$$[\nabla_x f]_k = \frac{df}{dx_k} = \begin{cases} 2Q_{kk} x_k & i = j = k \\ Q_{kj} x_j & i = k, i \neq j \\ Q_{ik} x_i & j = k, j \neq i \end{cases} \tag{11}$$

Therefore

$$\nabla_x f = (Q + Q^T)x \tag{12}$$

If Q is symmetric, then this equals $2Qx$.

4.1.2 Application to Least Squares

Let $f(w) = \|y - Xw\|_2^2$. Then the least squares problem is

$$\hat{w} = \arg\min_w f(w) \tag{13}$$

We can expand $f(w)$ as

$$\begin{aligned} f(w) &= (y - Xw)^T(y - Xw) \\ &= y^T y - y^T Xw - w^T X^T y + w^T X^T Xw \\ &= y^T y - 2w^T X^T y + w^T X^T Xw \end{aligned}$$

Then

$$\nabla_w f(w) = -2X^T y + 2X^T Xw$$

At an optimum we have that $\hat{w}$ solves $X^T y = X^T Xw$. Then if $(X^T X)^{-1}$ exists, we have that

$$\hat{w} = (X^T X)^{-1} X^T y \tag{14}$$

Theorem 1. *(Sufficient Condition for Existence/Uniqueness of LS Solution) If the columns of X are linearly independent, then $X^T X$ is non-singular, and there exists a unique least squares solution $\hat{w} = (X^T X)^{-1} X^T y$.*

4.2 Positive Definite Matrices

Definition 3 (Positive Definite, pd)**.** *A matrix Q ($n \times n$) is positive definite (written $Q \succ 0$) if $x^T Q x > 0$ for all $x \in \mathbb{R}^n$, $x \neq 0$.*

Definition 4 (Positive Semi-Definite, psd)**.** *A matrix Q ($n \times n$) is positive semi-definite (written $Q \succeq 0$) if $x^T Q x \geq 0$ for all $x \in \mathbb{R}^n$, $x \neq 0$.*

Properties of Positive Definite matrices:

1. If $P \succ 0$ and $Q \succ 0$, then $P + Q \succ 0$.
2. If $P \succ 0$ and $\alpha > 0$, then $\alpha P \succ 0$.
3. For any matrix A, $A^T A \succeq 0$ and $AA^T \succeq 0$. Further, if the columns of A are linearly independent, then $A^T A \succ 0$.

4. If $A \succ 0$, then A^{-1} exists.

5. Notation: $A \succ B$ means $A - B \succ 0$.

Example 5. Let

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \tag{15}$$

Then

$$X^T X = \begin{pmatrix} 3 & 3 \\ 3 & 3 \end{pmatrix} \tag{16}$$

Consider the vector $a = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Then $a^T X^T X a = 0$. Therefore $X^T X$ is not positive definite.

4.3 Subspaces

Definition 5. *(Subspace) A set of points $S \subseteq \mathbb{R}^n$ is a subspace if*

1. *$0 \in S$ (S contains the origin)*
2. *If $x, y \in S$, then $x + y \in S$*
3. *If $x \in S$, $\alpha \in \mathbb{R}$, then $\alpha x \in S$.*

4.4 Least Squares with Orthonormal Basis for Subspace

5 Least Squares Classification

We are given a training sample $\{x_i, y_i\}_{i=1}^n$, $x_i \in \mathbb{R}^p$ and $y \in \mathbb{R}$ (or $y \in \{+1, -1\}$).

Definition 6. *(Linear Predictor) We have a linear predictor if each label is a linear combination of the features i.e. we can find weights $\{w_i\}_{i=1}^p$ such that*

$$y_i = w_1 x_{i1} + w_2 x_{i2} + \dots w_p x_{ip} \tag{17}$$

In words, this says the label for observation i is a linear combination of the features for example i.

The steps to complete least squares classification in this environment are as follows:

1. Build a data matrix or feature matrix and label vector

$$X = \begin{bmatrix} \text{---} x_1^T \text{---} \\ \text{---} x_2^T \text{---} \\ \vdots \\ \text{---} x_n^T \text{---} \end{bmatrix} = \begin{bmatrix} x_1^T & 1 \\ x_2^T & 1 \\ \vdots & \\ x_n^T & 1 \end{bmatrix} \in \mathbb{R}^{n \times p}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{18}$$

The linear model is then $\hat{y} = Xw$.

2. Solve a least squares optimization problem

$$\hat{w} = \arg\min_{w} \|y - Xw\|_2^2 = \arg\min_{w} \sum_{i=1}^{n} (y_i - x_i^T w)^2 \tag{19}$$

(this last equality makes it clear that we are minimizing the sum of squared residuals). If the columns of X are linearly independent, then $X^T X$ is positive definite. Therefore $X^T X$ is invertible. In sum, if $X^T X$ is positive definite, then there exists a unique LS solution

$$\hat{w} = (X^T X)^{-1} X^T y \tag{20}$$

The predicted labels are

$$\begin{aligned} \hat{y} &= Xw \\ &= X(X^T X)^{-1} X^T y \end{aligned}$$

5.1 Regularization