

CS195-5, Lecture 4: Derivations and notes  
by Greg Shakhnarovich gregory@cs

## Slide 5

What is the relationship between the Euclidean distance  $\|\mathbf{x} - \mu\|$  and the argument of the exponent? Let us start with writing out the expression for the distance between two vectors  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ :

$$\|\mathbf{x} - \mathbf{y}\| \triangleq \sqrt{\sum_{j=1}^d (x_j - y_j)^2} \quad (1)$$

This is the familiar metric we normally use in Euclidean spaces; it measures a length of the shortest path between  $\mathbf{x}$  and  $\mathbf{y}$ . A closely related quantity is the Euclidean *norm* of a vector  $\mathbf{x}$ ,

$$\|\mathbf{x}\|^2 \triangleq \sum_{j=1}^d x_j^2.$$

This is of course equal to  $\mathbf{x}^T \mathbf{x}$ .

So, the norm of  $\mathbf{x} - \mu$  is equal to the squared distance of between  $\mathbf{x}$  and  $\mu$ , and also to

$$(\mathbf{x} - \mu)^T (\mathbf{x} - \mu).$$

What happens if we put a matrix  $\mathbf{A}$  “in the middle” of the dot product,  $(\mathbf{x} - \mu)^T \mathbf{A} (\mathbf{x} - \mu)$ ? The simplest case is the identity  $\mathbf{A} = \mathbf{I}$  - it of course does not change anything. Now suppose that  $\mathbf{A} = a\mathbf{I}$ , i.e. a diagonal matrix where all elements are zero except for the main diagonal which contains  $a$ . Then,

$$(\mathbf{x} - \mu)^T a\mathbf{I} (\mathbf{x} - \mu) = [x_1 - \mu_1, \dots, x_d - \mu_d] \begin{bmatrix} a & 0 & \dots & 0 \\ 0 & a & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & a & \dots \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ \vdots \\ x_d - \mu_d \end{bmatrix}$$

The first product yields  $[a(x_1 - \mu_1), \dots, a(x_d - \mu_d)]$  and then the second yields

$$a \sum_{j=1}^d (x_j - \mu_j)^2,$$

that is,  $\mathbf{A}$  simply *scaled* the distance by a factor of  $a$ .

A slightly more general case is  $\mathbf{A} = \text{diag}(a_1, \dots, a_d)$ , i.e a diagonal matrix with different elements on the main diagonal. It is easy to see that

$$(\mathbf{x} - \mu)^T \mathbf{A} (\mathbf{x} - \mu) = \sum_{j=1}^d a_j (x_j - \mu_j)^2$$

This is a more complex scaling operation. In terms of distances, one interpretation is that  $a_j$  specifies the “cost” of traveling in the direction parallel to the axis  $x_j$ .

We will defer the discussion of more general cases for later.

## Slide 9

Here is the derivation for the mean and covariance of  $\hat{\mathbf{w}}$ , given the training data  $\mathbf{X}$ . We will start dropping the part of the notation that specifies the distribution with respect to which the expectations are taken whenever that’s obvious.

$$E[\hat{\mathbf{w}}|\mathbf{X}] = E[\mathbf{w}^* + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \nu | \mathbf{X}] \quad (2)$$

$$= \mathbf{w}^* + E[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \nu | \mathbf{X}] \quad (3)$$

$$= \mathbf{w}^* + \mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E[\nu | \mathbf{X}] \quad (4)$$

$$= \mathbf{w}^* + \mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E[\nu] \quad (5)$$

$$= \mathbf{w}^* + \mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{0} \quad (6)$$

$$= \mathbf{w}^*. \quad (7)$$

Justification for the steps above:

(3)  $\mathbf{w}^*$  is independent of a particular  $\mathbf{X}$ ; it’s the optimal linear regressor for the *model*, not the data. Therefore, it’s a constant with respect to  $\mathbf{X}$  and we can take it out of the expectation.

(4)  $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}$  is of course also a constant w.r.t.  $\mathbf{X}$ .

(5) since in our model the noise is independent of the inputs,  $p(\nu | \mathbf{X}) = p(\nu)$  and we can remove conditioning on  $\mathbf{X}$ .

(6) assumption that the noise is white (zero-mean).

Recall that the covariance of two random variables  $a$  and  $b$  is defined as

$$\text{Cov}_{a,b} \triangleq E_{p(a,b)} [(a - \mu_a)(b - \mu_a)], \quad (8)$$

where  $\mu_a = E_{p(a)} [a]$  and  $\mu_b = E_{p(b)} [b]$  are the *marginal* means of the corresponding random variables. Intuitively, this means how  $a$  and  $b$  “co-vary” (i.e., if  $\text{Cov}_{a,b}$  is positive it means that they tend, in probability, to deviate to the same direction from their means.) A related quantity is correlation,

$$\text{cor } a, b \triangleq \frac{\text{Cov}_{a,b}}{\sigma_a \sigma_b},$$

where  $\sigma_a$  denotes the standard deviation (square root of variance) of  $a$ . Correlation measures the amount of *linear* relationship between the two variables. Note that both covariance and correlation are *symmetric*, in the sense that  $\text{Cov}_{a,b} = \text{Cov}_{b,a}$ .

The covariance matrix  $\text{Cov}_{\mathbf{z}}$  of a random variable  $\mathbf{z} \in \mathbb{R}^d$  is a generalization of this concept. The  $(i, j)$  entry in this matrix is the covariance of  $z_i$  and  $z_j$ ; the diagonal elements are therefore just the variances of  $z_i, i = 1, \dots, d$  (sometimes this matrix is called the *variance-covariance* matrix):

$$\text{Cov}_{\mathbf{z}} \triangleq \begin{bmatrix} \sigma_{z_1}^2 & \text{Cov}_{z_1, z_2} & \dots & \text{Cov}_{z_1, z_d} \\ \text{Cov}_{z_2, z_1} & \sigma_{z_2}^2 & \dots & \text{Cov}_{z_2, z_d} \\ \vdots & & \ddots & \vdots \\ \text{Cov}_{z_d, z_1} & \text{Cov}_{z_d, z_2} & \dots & \sigma_{z_d}^2 \end{bmatrix}.$$

From this definition a few properties follow immediately:

1.  $\text{Cov}_{\mathbf{z}}$  is square and symmetric.
2. The elements on the main diagonal are always non-negative (and can be zero only if the corresponding  $z_i$  is constant).

It is easy now to show that under the above definition,

$$\text{Cov}_{\mathbf{z}} = E [(\mathbf{z} - \mu_{\mathbf{z}})(\mathbf{z} - \mu_{\mathbf{z}})^T]$$

A quick “dimension sanity check”: this is an *outer product*, of a  $d \times 1$  column vector by a  $1 \times d$  row vector, yielding a  $d \times d$  matrix.

Here is an often useful particular case. Suppose we have a random vector  $\mathbf{z}$  the components  $z_j$  of which are known to be statistically independent. Then, by definition of covariance, the elements off the diagonal are zero, and the covariance matrix is  $\text{Cov}_{\mathbf{z}} = \text{diag}(\sigma_{z_1}^2, \dots, \sigma_{z_d}^2)$ . If the components of  $\mathbf{z}$  are *identically* distributed, so that they all have the same variance  $\sigma^2$ , the covariance matrix is just  $\sigma^2 \mathbf{I}$ . Note that under the statistical model we have assumed, this is exactly the case for our random noise vector  $\nu$ .

Now let's get back to deriving the covariance of  $\hat{\mathbf{w}}$ . Recall that under our current assumptions,

$$\hat{\mathbf{w}} = \mathbf{w}^* + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X} \nu$$

We can use the result we have just derived that  $E[\hat{\mathbf{w}}] = \mathbf{w}^*$

$$\text{Cov}_{\mathbf{w}} E[\hat{\mathbf{w}}|\mathbf{X}] = E[(\hat{\mathbf{w}} - \mathbf{w}^*)(\hat{\mathbf{w}} - \mathbf{w}^*)^T | \mathbf{X}] \quad (9)$$

$$= E\left[\left((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X} \nu\right) \left((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X} \nu\right)^T \middle| \mathbf{X}\right] \quad (10)$$

$$= E\left[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X} \nu \nu^T \mathbf{X}^T (\mathbf{X}^T \mathbf{X})^{-1} \middle| \mathbf{X}\right] \quad (11)$$

$$= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X} E[\nu \nu^T \mathbf{X}^T (\mathbf{X}^T \mathbf{X})^{-1} | \mathbf{X}] \quad (12)$$

$$= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X} E[\nu \nu^T] \mathbf{X}^T (\mathbf{X}^T \mathbf{X})^{-1} \quad (13)$$

$$= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X} \sigma^2 \mathbf{I} (\mathbf{X}^T \mathbf{X})^{-1} \quad (14)$$

$$= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \quad (15)$$

$$= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}. \quad (16)$$

You should at this point be able to justify all the steps above. Look for similar steps in the derivation for the mean, and also go over the matrix tutorial by Sam Roweis, posted under Lecture 2 material on the website.