Introduction to Machine Learning **Fall Semester**

Recitation 11

Lecturer: Regev Schweiger *Scribe: Regev Schweiger*

11.1 Kernel Ridge Regression

We now take on the task of kernel-izing ridge regression. Let $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^d$, and $\mathbf{y} \in \mathbb{R}^m$. Recall that ridge regression solves the following problem:

$$\arg\min_{\mathbf{a}\in\mathbb{R}^d} \|\mathbf{y} - X\mathbf{a}\|^2 + \lambda\|\mathbf{a}\|^2$$

where λ is penalty coefficient. Equating the gradient to 0 result in the solution we have seen in class:

$$\hat{\mathbf{a}} = (X^T X + \lambda I_d)^{-1} X^T \mathbf{y}$$

Note that

$$(X^T X + \lambda I_d)X^T = X^T X X^T + \lambda X^T = X^T(XX^T + \lambda I_n)$$

Multiplying $(X^T X + \lambda I_d)^{-1}$ at the left and $(XX^T + \lambda I_n)^{-1}$ at the right, we get

$$X^T(XX^T + \lambda I_n)^{-1} = (X^T X + \lambda I_d)^{-1} X^T$$

Therefore, the optimal solution is equivalently,

$$\hat{\mathbf{a}} = X^T(XX^T + \lambda I_n)^{-1}\mathbf{y}$$

Given a new point $\mathbf{x}$, our regression estimate will be

$$\mathbf{x}^T\hat{\mathbf{a}} = \mathbf{x}^T X^T(XX^T + \lambda I_n)^{-1}\mathbf{y}$$

We would now like to embed our points to a space H, with $\mathbf{x}_i \to \phi(\mathbf{x}_i)$, and perform ridge regression after the transformation. It is easy to see that, given the formulation above, we can replace all the expressions involving X with kernel expressions. First define $\bar{K}$ as the matrix for which $\bar{K}_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$. Similarly define $\mathbf{k}$ as the vector for which $\mathbf{k}_i = \phi(\mathbf{x})^T\phi(\mathbf{x}_i) = K(\mathbf{x}, \mathbf{x}_i)$. Thus, given a new point $\mathbf{x}$, our regression estimate will be

$$\phi(\mathbf{x})^T\hat{\mathbf{a}} = \mathbf{k}^T(\bar{K} + \lambda I_n)^{-1}\mathbf{y}$$

Note that, as usual, we cannot write down $\hat{\mathbf{a}}$ explicitly, but we can apply it to the transformation of new points.

11.2 PCA as maximizing variance

We have seen how the PCA algorithm can be derived in the context of minimizing the reconstruction error. More formally, assume we have a set of input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$, where $\mathbf{x}_i \in \mathbb{R}^d$. Denote the principal components by the columns of V, as $\mathbf{v}_1, \ldots, \mathbf{v}_r$; the orthonormality constraints imply that $V^T V = I$. The PCA problem was:

$$V = \arg \min_{V \in \mathbb{R}^{d \times r}} \sum_{i=1}^{m} \|\mathbf{x}_i - VV^T \mathbf{x}_i\|^2. \tag{11.1}$$

We now consider another possible criterion. Let's assume $r = 1$, that is, we would like to find the "best" line in some sense. One intuitive criterion is the line, which if we project all points on, will give maximal empirical variance. The empirical variance of a set of measurements, $a_1, \ldots, a_m$, is

$$\frac{1}{m} \sum_{i=1}^{m} (a_i - \frac{1}{m} \sum_{j=1}^{m} a_j)^2$$

Assume without loss of generality that the data points are centered at zero, that is

$$\sum_{i=1}^{m} \mathbf{x}_i = 0$$

If that is not the case, we mean-center the data. Therefore, it is easy to say that $\sum_{i=1}^{m} \mathbf{v}^T \mathbf{x}_i = 0$ for each $\mathbf{v}$. Therefore, the empirical variance of the set of projection is simply the mean of squares. Therefore, the criterion we like for the first direction is:

$$\mathbf{v}_1 = \arg \max_{\|\mathbf{v}\|=1} \frac{1}{m} \sum_{i=1}^{m} (\mathbf{v}^T \mathbf{x}_i)^2$$

For the next direction, we would like to capture the variance on directions we have not yet seen. Formally, we would like directions orthogonal to previous directions. Assume we found already $\mathbf{v}_1, \ldots, \mathbf{v}_{r-1}$. Then, the r-th direction is:

$$\mathbf{v}_r = \text{argmax}_{\|\mathbf{v}\|=1, \mathbf{v} \perp \mathbf{v}_1, \ldots, \mathbf{v}_{r-1}} \frac{1}{m} \sum_{i=1}^{m} (\mathbf{v}^T \mathbf{x}_i)^2$$

We can instead formulate that to find all r directions together, to get:

$$\text{argmax}_{V \in \mathbb{R}^{d \times r}, V^T V = I} \sum_{j=1}^{r} \frac{1}{m} \sum_{i=1}^{m} (\mathbf{v}_j^T \mathbf{x}_i)^2$$

It is easy to see that the optimization function is:

$$\sum_{j=1}^{r} \frac{1}{m} \sum_{i=1}^{m} (\mathbf{v}_j^T \mathbf{x}_i)^2 = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{r} (\mathbf{v}_j^T \mathbf{x}_i)^2 = \frac{1}{m} \sum_{i=1}^{m} \|V^T \mathbf{x}_i\|^2$$

To summarize, a sensible criterion for dimensionality reduction would be to choose V so that the variance of projections is maximized, i.e., intuitively the structure of the data is preserved as much as possible:

$$\text{argmax}_{V \in \mathbb{R}^{d \times r}, V^T V = I} \sum_{i=1}^{m} \|V^T \mathbf{x}_i\|^2. \tag{11.2}$$

We note, however, the following equality, based on Pythagoras:

$$\|\mathbf{x}_i\|^2 = \|VV^T \mathbf{x}_i\|^2 + \|\mathbf{x}_i - VV^T \mathbf{x}_i\|^2.$$

And it is easy to see that $\|VV^T \mathbf{x}_i\|^2 = \|V^T \mathbf{x}_i\|^2$ due to the orthonormality of V. Since $\|\mathbf{x}_i\|$ does not depend on V, we see that **minimizing the reconstruction error is equivalent to maximizing the variance**. The goal in principal component analysis (PCA) is therefore to minimize the reconstruction error (see Equation 11.1), and to maximize the projected variance (Equation 11.2).

Eigenvalues. An important observation is the following. We know that the solution of PCA is the eigenvectors of the empirical covariance matrix. What are the eigenvalues? The variance maximization criterion gives an intuitive interpretation. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $C = \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^T$; i.e., $C\mathbf{v}_j = \lambda_i \mathbf{v}_j$. We seeked to maximize $\frac{1}{m} \sum_{i=1}^{m} (\mathbf{v}^T \mathbf{x}_i)^2$. Plugging in $\mathbf{v} = \mathbf{v}_j$, we get

$$\frac{1}{m} \sum_{i=1}^{m} (\mathbf{v}_j^T \mathbf{x}_i)^2 = \frac{1}{m} \mathbf{v}_j^T \left(\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^T \right) \mathbf{v}_j = \mathbf{v}_j^T C \mathbf{v}_j = \lambda_j$$

That is, λ_j, the j-th eigenvalue, is the empirical variance of the projection on the j-th principal axis.

11.3 PCA example

11.3.1 Background

The DNA in our cells contains long chains of four chemical building blocks – adenine, thymine, cytosine, and guanine, abbreviated A, T, C, and G. More than 6 billion of these

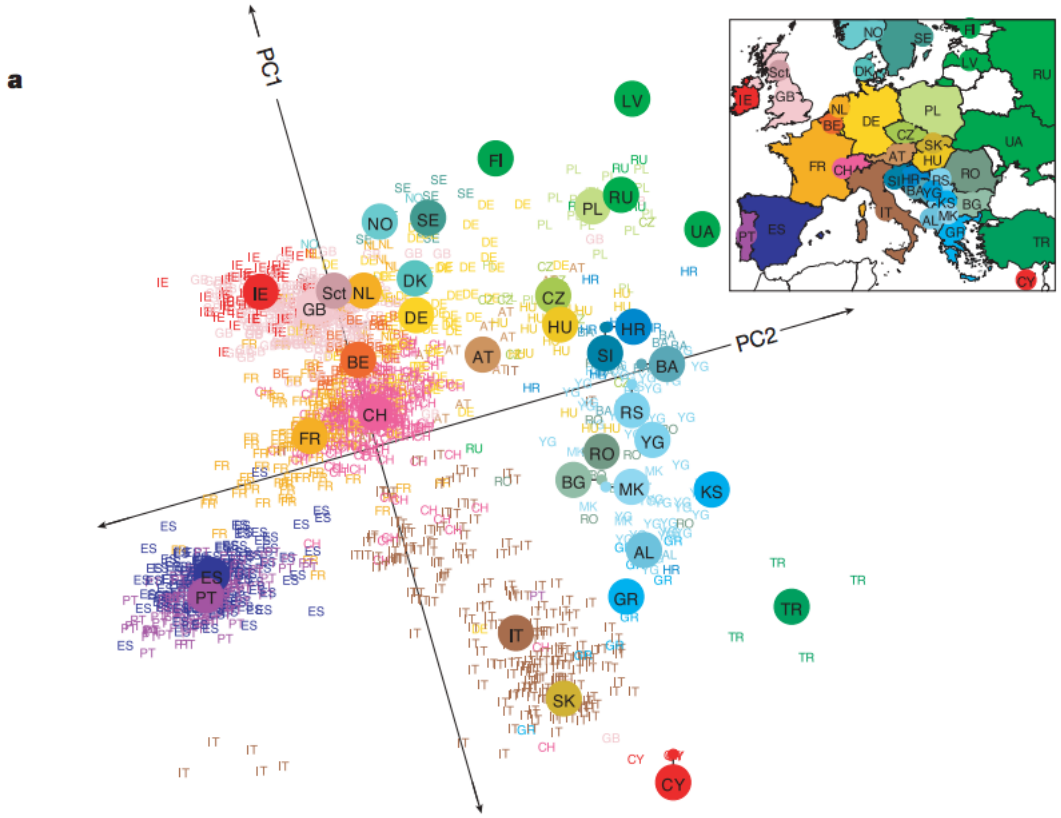
chemical bases, strung together in 23 pairs of chromosomes, exist in a human cell. These genetic sequences contain information that influences our physical traits, our likelihood of suffering from disease, and the responses of our bodies to substances that we encounter in the environment.

The genetic sequences of different people are remarkably similar. When the chromosomes of two humans are compared, their DNA sequences can be identical for hundreds of bases. But at about one in every 1,200 bases, on average, the sequences will differ. Differences in individual bases are by far the most common type of genetic variation. One person might have an A at that location, while another person has a G. These genetic differences are known as single nucleotide polymorphisms, or SNPs (pronounced "snips"). There are approximately 10 million SNPs estimated to occur commonly in the human genome. Each distinct "spelling" of a chromosomal region is called an allele, and a collection of alleles in a person's chromosomes is known as a genotype.

In the most common case, there are only two alleles for all population at each SNP position. Data describing the genotype data for individuals, often does not specify the bases explicitly. Instead, one allele (per position) is selected as a reference allele. Then, at that position, the number of non-reference alleles is presented: 0 if both alleles in that position, in the chromosome pair, were identical to the reference allele for that position; 1 if only one of them was the reference allele; and 2 if neither were the reference alleles.

11.3.2 Novembre et al., 2008

In the work of *Novembre et al. 2008, Nature*, 3,192 European individuals were genotyped at 500,568 positions (some details are omitted for simplicity). They applied PCA with $r = 2$ and presented the projections of all genomes on these two principal axes:



Each individuals is denoted by colored two-letters, denoting their country of origin. It can be seen that the projections reflect the geography of Europe well.