LECTURE 29

I. The Least Squared Approximation:
   Its cognitive value in data mining.

II. Examples of Least Squares Problems
   (i) Algebraic solutions
   (ii) Geometric solutions

III. The Fundamental Theorem:

   Orthogonal Projection = Optimal Projection.
I. THE LEAST SQUARES APPROXIMATION:

Of What Value Is it? Why?

Least squares approximations condense a dataset into a cognitively manageable, quantitative form whose value arises as a means to an end in data mining.

The cognitive elements of this condensation process vary from one experiment/observation to another, even though the experiments/observations are identically designed/executed (within the context of technological limitation).

These condensations [i.e., the to-be-calculated least squares approximations] have distinct but commensurable identities. In probability theory, they are expressed quantitatively as "random" (or "stochastic") variables. In this theory, one defines a probability function on this domain of random variables, and thereby arrives at potentially valuable knowledge about the aggregate of least squares approximations.
in the best way.

\[ w_k^* = b_0 + b_1 x_1 + b_2 x_2 \]

Here "best" means

\[ \| w_k^* - \hat{w} \|^2 \geq \| w_k^* - \hat{w} \|^2 \]

for all \( \hat{w} = \{ b_0 + b_1 x_i + b_2 x_i \}_{i=1}^m \).

i.e. for all

\[
\begin{bmatrix}
  b_0 \\
  b_1 \\
  b_2
\end{bmatrix} \in \mathbb{R}^3
\]

and then by projecting \( \hat{w}_k \) onto the subspace

\[ W_3 = \text{span}(\hat{u}_1, \hat{u}_2, \hat{u}_3) \subset V = \mathbb{R}^3 \]

determine the optimal (least-squares) solution

\[
\begin{bmatrix}
  b_0^*_k \\
  b_1^*_k \\
  b_2^*_k
\end{bmatrix}
\]

corresponding to the \( k \)-th given dataset \((x, \hat{w})\).

\[ w(x_i) = w_k^* \quad i = 1, 2, \ldots, m \]

**SOLUTION:** Rewrite this system in the form

\[
\begin{bmatrix}
  1 & x_1 & x_1^2 \\
  1 & x_2 & x_2^2 \\
  \vdots & \vdots & \vdots \\
  1 & x_m & x_m^2
\end{bmatrix}
\begin{bmatrix}
  b_0 \\
  b_1 \\
  b_2
\end{bmatrix} =
\begin{bmatrix}
  w_1^* \\
  w_2^* \\
  \vdots \\
  w_m^*
\end{bmatrix}
\]

where \( \hat{u}_1, \hat{u}_2, \hat{u}_3 \in W_3 \) and then project \( \hat{u}_k \) onto the subspace

\[ W_3 = \text{span}(\hat{u}_1, \hat{u}_2, \hat{u}_3) \subset V = \mathbb{R}^3 \]

determine the optimal (least-squares) solution

\[
\begin{bmatrix}
  b_0^*_k \\
  b_1^*_k \\
  b_2^*_k
\end{bmatrix}
\]

corresponding to the \( k \)-th given dataset \((x, \hat{w})\).
\[ W_N = \text{Sp}(\mathbf{u}_0, \mathbf{u}_N) \quad (N=3 \text{ in the preceding example}) \]

b) Geometrically, the task is to find the best vector

\[ \mathbf{w}^{*N} = b_0 \mathbf{u}_0 + b_1 \mathbf{u}_1 + b_2 \mathbf{u}_2 \]

Comment 1: Instead of basing the optimization on a 3-parameter polynomial, an alternative least squares is possible namely use 3-term combinations of some linear independent set of 3 functions.

Comment 2:

Familiarity with the properties of the shortest line between a point and a plane suggests that the error vector

\[ \mathbf{e} = \mathbf{w}^{*N} \]

be perpendicular to every vector in the subspace \( W_N \), i.e. \( \mathbf{n}^T \mathbf{w}^{*N} = 0 \). This observation is made precise by the following fundamental
**Fundamental Theorem:**

Given:

1. $V$ is an inner product space.
2. $W_n \subseteq V$ is a subspace of $V$.
3. $\overrightarrow{v} \in V$ any vector in $V$.

Conclusion:

There exists a unique $\overrightarrow{w}^{**} \in W_n$ such that

$$\langle \overrightarrow{v} - \overrightarrow{w}^{**} | \overrightarrow{w}^{**} \rangle = 0 \quad \forall \overrightarrow{w} \in W_n$$

if and only if ($\Rightarrow$)

There exists a unique $\overrightarrow{w}^{**} \in W_n$ such that

$$\| \overrightarrow{v} - \overrightarrow{w}^{**} \| < \| \overrightarrow{v} - \overrightarrow{w} \| \quad \forall \overrightarrow{w} \in W_n$$

or equivalently

$$\varphi_k(\overrightarrow{w}) = \| \overrightarrow{v} - \overrightarrow{w} \|^2 \text{ has a unique minimum at } \overrightarrow{w} = \overrightarrow{w}^{**}.$$
proof: (\(\Leftrightarrow\)) (in 4 steps)

**Step I** Introduce an o.n. basis for \(W_N\)

\[ \vec{w} = c_1 e_1 + \ldots + c_N e_N, \text{ where } c_k = a_k + i b_k \]
and consider the error \(e(x)\)

and for typographical reasons suppress the superscript \(k\) and the overarrow \(\vec{\cdot}\):

\[ e_k^2 = \| e_k \|^2 = \sum_{k=1}^{N} c_k \overline{c_k} = 1 \]

\[ = \sum_{k=1}^{N} \langle e_k, e_k \rangle = \sum_{k=1}^{N} (a_k^2 + b_k^2) \]

\[ = \sum_{k=1}^{N} \langle e_k, e_k \rangle = \sum_{k=1}^{N} (\overline{a_k} a_k + \overline{b_k} b_k) \]

**Step II**

\[ e_k^2 = 2 a_k (\overline{a_k} e_k + \overline{b_k} e_k) = 0 \]

Hence: \( a_k = \frac{1}{2} \langle w_k e_k \rangle + \frac{1}{2} \langle \bar{w}_k e_k \rangle \)

\[ 2 b_k = 2 b_k - i (\overline{a_k} e_k + \overline{b_k} e_k) = 0 \]

Hence: \( b_k = \frac{1}{2} \langle w_k e_k \rangle - \frac{\overline{a_k}}{2} \langle \bar{w}_k e_k \rangle \)

Thus: \( a_k \overline{b_k} = c_k = \langle \bar{w}_k | e_k \rangle \) - unique vector

\[ \bar{w}_k^* = \sum_{k=1}^{N} \langle \bar{w}_k | e_k \rangle e_k \] in \(W_N\)

**Step III**

Using this critical vector one finds

\[ \langle \bar{w}_k^* | \bar{w}_k \rangle = \langle \bar{w}_k^* | \bar{w}_k \rangle = \sum_{k=}^{N} c_k (\overline{\bar{w}_k^* e_k}) = \sum_{k=}^{N} c_k (\overline{\bar{w}_k^* e_k}) = 0 \]

Thus: \( \bar{w}_k^* \) is \( \perp W_N \) indeed.

**Step IV**

\[ \frac{\partial^2 E^2}{\partial a_k^2} = 2 \langle w_k e_k \rangle, \frac{\partial^2 E^2}{\partial b_k^2} = -2 \rightarrow 0 \]

Thus \( E^2 \) is a minimum, i.e. \( \bar{w}_k^* \) is the shortest (= best) vector.

**Comment**

It is not necessary to use an o.n. basis for \(W_N\) in order to find the optimal.

This makes computations easy. This fact is expressed by the following theorem in LECTURE 30.