Linear algebra & the exponential family

Lloyd Elliott

October 25, 2010

Bilinear forms

Exponential family

Bilinear forms

Definition

A *bilinear form* is a map

$$f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$$

with the following properties:

- $f(x + y, z) = f(x, z) + f(y, z)$
- $f(z, x + y) = f(z, x) + f(z, y)$
- $f(cx, y) = f(x, cy) = cf(x, y)$

If f is a bilinear form then there is a matrix A such that

$$f(x, y) = x^T A y$$

Conversely, $f_A(x, y) = x^T A y$ is a bilinear form.

A bilinear form is *symmetric* if $f(x, y) = f(y, x)$.
This happens if and only if f_A is symmetric.
A bilinear form is *positive definite* if $f(x, x) \geq 0$ for all x with equality only if $x = 0$. A bilinear form is *positive semi-definite* if $f(x, x) \geq 0$ but equality may hold more generally. A matrix is positive (semi-)definite if f_A is positive (semi-)definite.

Example

$$f_I(x, y) = \langle x, y \rangle$$

Definition

Let A be a matrix. Suppose

$$Av = \lambda v$$

Then v is called an *eigenvector* for A and λ its corresponding *eigenvalue*.

Theorem (Spectral theorem)

Suppose Q is a symmetric matrix. There exists an orthanormal basis $v_1, \ldots, v_n$ of $\mathbb{R}^n$ such that $v_1, \ldots, v_n$ are eigenvectors for A.
Suppose Q is a positive semi-definite matrix. Then the corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$ are ≥ 0.
Suppose Q is a positive definite matrix. Then $\lambda_1, \ldots, \lambda_n$ are strictly positive.

Theorem (Singular value decomposition)

If A is an $m \times n$ matrix then there is an orthanormal basis $v_1, \ldots, v_n$ of $\mathbb{R}^n$ and an orthanormal basis $w_1, \ldots, w_m$ of $\mathbb{R}^m$ and $\sigma_1 \geq \ldots \geq \sigma_r > 0$ such that

$$Av_i = \begin{cases} \sigma_i w_i & \text{if } 1 \leq i \leq r \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, the σ_i are unique up to permutation.

The vectors $\sigma_1, \ldots, \sigma_r$ are referred to as the singular values for A.

Proof.

Let $S = A^T A$. By the spectral theorem, S has eigenvalues $\lambda_1, \ldots, \lambda_r > 0$ and eigenvectors $v_1, \ldots, v_n$ with $\lambda_i = 0$ for $r < i \leq n$. Let $\sigma_i = \sqrt{\lambda_i}$ and let $w_i = \frac{1}{\sigma_i} Av_i$ for $1 \leq i \leq r$. Extend $w_1, \ldots, w_r$ to an orthanormal basis of $\mathbb{R}^m$. Note that since $Av_i = 0$ for $i > r$, the vectors $w_{r+1}, \ldots, w_m$ do not contribute to the representation. □

Corollary

If A is an $m \times n$ matrix then there is an orthogonal $m \times m$ matrix U and an orthogonal $n \times n$ matrix V such that:

$$A = U\Sigma V^T$$

where Σ is the $m \times n$ matrix such that

$$\Sigma_{ij} = \begin{cases} \sigma_i & \text{if } i = j \leq r \\ 0 & \text{otherwise.} \end{cases}$$

This means that every matrix is a rotation and then some scaling followed by another rotation. Since the rows of Σ are zero beyond the r-th row, we may drop all but the first r rows of Σ and V and still have equality. This reduces the dimensionality of the representation V of A. And $U\Sigma$ describes how to move from the reduced representation back to A.

Theorem

Define $\hat{\Sigma}^{(k)}$ *for* $k \leq r$ *to be:*

$$\hat{\Sigma}^{(k)} = \begin{cases} \sigma_i & \text{if } i = j \leq k \\ 0 & \text{otherwise.} \end{cases}$$

Then, $\hat{\Sigma}^{(k)} = \text{argmin}_{B|\ \text{rank}B=k} \sum_{ij} |A_{ij} - B_{ij}|^2$.

This theorem means that if we drop all but the first k rows then we recover the best mean squared error representation of A possible with a rank-k representation.

Exponential family

Let X be a random vector. The pdf of X is an *exponential family* distribution if it is of the form:

$$p(x|\nu) = g(\nu)f(x)\exp(\nu^T S(x)).$$

This is called a cannonical representation because there is no function of ν in the exponent. The canonical form of an exponential family is unique up to the choice of the function $S(x)$.

- $g(\theta)$ is called the inverse partition function.
- $S(x)$ is called the vector of sufficient statistics.
- ν is called the vector of natural parameters.
- $f(x)$ is called the base measure.

Theorem

$$\mathbb{E}[S_j(x)] = -\frac{\partial}{\partial \nu_j} \log(\nu).$$

Proof.

$$\begin{aligned}
\mathbb{E}[S_j(x)] &= \int_{-\infty}^{\infty} S_j(x) g(\nu) f(x) \exp(\nu^T S(x)) \mathrm{d}x, \\
&= g(\nu) \int_{-\infty}^{\infty} S_j(x) f(x) \exp\left(\sum_i \nu_i S_i(x)\right) \mathrm{d}x, \\
&= g(\nu) \int_{-\infty}^{\infty} \frac{\partial}{\partial \nu_j} f(x) \exp\left(\sum_i \nu_i S_i(x)\right) \mathrm{d}x, \\
&= g(\nu) \frac{\partial}{\partial \nu_j} \int_{-\infty}^{\infty} f(x) \exp(\nu^T S(x)) \mathrm{d}x.
\end{aligned}$$

continued

□

Proof.

$$= g(\nu)\frac{\partial}{\partial \nu_j} 1/g(\nu) \tag{1}$$

$$= \frac{\partial}{\partial \nu_j} \log(g(\nu)) \tag{2}$$

As $\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$, $\int_{-\infty}^{\infty} f(x)\exp(\nu^T S(x)) = 1/g(\nu)$ and this is where $1/g(\nu)$ comes from in step (1). And (2) is from $\frac{\mathrm{d}}{\mathrm{dx}} \log(f(x)) = f(x)/f'(x)$. □