Math Modeling, Week 14

Hilbert space

Generalizes Euclidean space to possibly infinite dimensions
- Useful for studying spaces of functions, sequences, etc.

Real vector space
- Addition: $f + g \in \mathcal{H}$ for all $f, g \in \mathcal{H}$
- Scalar multiplication: $rf \in \mathcal{H}$ for all $f \in \mathcal{H}$ and $r \in \mathbb{R}$ (or $r \in \mathbb{C}$ for complex vector space)

Inner product
- Generalizes dot product
- $\langle f, g \rangle \in \mathbb{R}$ for all $f, g \in \mathcal{H}$
- Symmetric: $\langle f, g \rangle = \langle g, f \rangle$
- Bilinear: $\langle f_1 + f_2, g \rangle = \langle f_1, g \rangle + \langle f_2, g \rangle$ and $\langle rf, g \rangle = r\langle f, g \rangle$
- Positive definite: $\langle f, f \rangle \geq 0$ and $\langle f, f \rangle = 0 \rightarrow f = 0$

Orthonormal basis
- Set of elements $\{f_i\} \subset \mathcal{H}$, possibly infinite
- Orthonormal: $\langle f_i, f_i \rangle = 1$ and $\langle f_i, f_j \rangle = 0$ for $i \neq j$
- Coordinate representation: any $f \in \mathcal{H}$ can be written (uniquely) as $f = \sum_i c_i f_i$, with $c_i = \langle f, f_i \rangle$
  - Note that $\sum_i c_i^2 = \langle f, f \rangle < \infty$
  - Explicit characterization of full space: $\mathcal{H} = \{\sum_i c_i f_i \,|\, \sum_i c_i^2 < \infty\}$
- Basis is not unique—can always rotate

Hilbert space of functions learnable on some stimulus set
- Expected reward, category membership probability, etc.
- Inner product will reflect stimulus representation, enabling link between feature and similarity models

Feature representations

Each stimulus represented as vector of feature values, $(\mathbf{x}_i)$
- Feature set could be finite, $i \in \{1, \ldots, m\}$, or infinite, $i \in \mathbb{N}$
- Assume features are linearly independent (non-redundant): $\forall \mathbf{x}(\sum c_i \mathbf{x}_i = 0) \leftrightarrow \forall i(c_i = 0)$

Let $\mathcal{H}$ be set of all linear functions $f: \mathcal{X} \rightarrow \mathbb{R}$ based on the features
- Linear combinations of feature functions: $f_i(\mathbf{x}) = \mathbf{x}_i$, $\mathcal{H} = \{\sum_i w_i f_i \,|\, \sum_i w_i^2 < \infty\}$
- Corresponds to all weight vectors: $\mathcal{H} = \{\langle w, \cdot \rangle \,|\, \sum_i w_i^2 < \infty\}$
- Also expressible as inner products with stimuli (guaranteed by non-redundancy assumption)
  - $f \in \mathcal{H} \rightarrow \exists \mathbf{x}^1, \ldots \in \mathcal{X}, c_i, \ldots \in \mathbb{R}\left(f = \sum_i c_i \langle \mathbf{x}^i, \cdot \rangle, \sum_i c_i^2 < \infty\right)$

Inner product on $\mathcal{H}$
- Features form orthonormal basis: $\langle f_i, f_i \rangle = 1$ and $\langle f_i, f_j \rangle = 0$ for $i \neq j$
- Mirrors inner product on stimuli: let $f_{\mathbf{x}} = \langle \mathbf{x}, \cdot \rangle = \sum_i \mathbf{x}_i f_i$, then $\langle f_{\mathbf{x}}, f_{\mathbf{x}'} \rangle = \langle \mathbf{x}, \mathbf{x}' \rangle$

Uniqueness
- Features fully determine the Hilbert space
- Hilbert space determines the features up to rotation
  - Rotation (change of coordinate system) has no effect on inner product

Similarity representation

Kernel or similarity function $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, with $k(x, y) = k(y, x)$

Each stimulus determines an evaluation function f_x, with $f_x(y) = k(x, y) = f_y(x)$

Let $\mathcal{H}$ be set of all functions learnable by summed similarity
- $\boldsymbol{\mathcal{H}} = \left\{\sum_i c_i f_{x^i} \,|\, x^i \in \mathcal{X}, \sum_i c_i^2 < \infty\right\}$
- Real vector space (addition, scalar multiplication)

Define inner product from kernel

$\langle f_x, f_{x'} \rangle = k(x, x')$

Extend by bilinearity: $\langle \sum_i c_i f_{x^i}, \sum_j c'_j f_{y^j} \rangle = \sum_{ij} c_i c'_j k(x^i, y^j)$

Implies reproducing property: $\langle f, f_x \rangle = f(x)$

Inner product is positive definite iff k is

Given stimuli $x^1, \ldots, x^n$, define kernel matrix K by $K_{ij} = k(x^i, x^j)$

Positive-definite kernel: K always positive semidefinite, $\forall b \in \mathbb{R}^n (b^{\mathrm{T}} K b \geq 0)$

Choose an orthonormal basis $\{f_i\}$ for $\mathcal{H}$

Guaranteed because $\mathcal{H}$ is a Hilbert space

Can treat as feature functions

Identify any stimulus x with vector of feature values, $\mathbf{x} = (f_i(x))$

<u>Kernel duality</u>

Dual representations

$k(x, x') = \langle \mathbf{x}, \mathbf{x}' \rangle$

Same space of learnable functions:

$f_x(y) = k(x, y) = \langle \mathbf{x}, \mathbf{y} \rangle = f_{\mathbf{x}}(\mathbf{y})$, so $f_x = f_{\mathbf{x}}$

$\mathcal{H} = \{\sum_i c_i f_{x^i} \,|\, x^i \in \mathcal{X}, \sum_i c_i^2 < \infty\} = \{\sum_i w_i f_i \,|\, \sum_i w_i^2 < \infty\}$

Equivalence of predictions

Feature model

Knowledge state: feature weights $\mathbf{w}_i$

Prediction for new stimulus x: $\langle \mathbf{w}, \mathbf{x} \rangle$

Exemplar model

Knowledge state: exemplars x^i and exemplar weights c_i

Prediction for new stimulus x: $\sum_i c_i k(x^i, x) = \langle \sum_i c_i \mathbf{x}^{\mathrm{i}}, \mathbf{x} \rangle$

Same predictions iff $\mathbf{w} = \sum_i c_i \mathbf{x}^i$

Equivalence of learning rules

Rescorla-Wagner: $\Delta \mathbf{w} = \varepsilon \delta \mathbf{x}_t$

Equivalent exemplar update rule

Let i_t be index of exemplar on trial t

$\Delta c_{i_t} = \varepsilon \delta$, i.e., update weight only for the presented stimulus

Exactly Pearce's (1987, 1994) configural learning model

<u>Regularization</u>

Discourage overfitting by modifying objective function with penalty for complexity of solution

Ridge regression

Extend ordinary least squares regression by penalizing sum of squared weights (L2 penalty)

Loss function: $\mathcal{L}(w) = \sum_i (\mathbf{x}^i w - y_i)^2 + \lambda \sum_j w_j^2 = \|Xw - Y\|^2 + \lambda \|w\|^2$

Optimal weights: $\widehat{w} = (X^{\mathrm{T}} X + \lambda I_m)^{-1} X^{\mathrm{T}} Y$

Bayesian interpretation

Gaussian prior on weight vector: $w \sim \mathcal{N}(0, \sigma_\eta^2 I_m)$

Data generated as in OLS regression: $X \sim \mathcal{N}(Xw, \sigma_\varepsilon^2 I_n)$

Posterior: $p(w|X, Y) \propto \exp\left(-\frac{1}{2}\left[\sigma_\varepsilon^{-2} \|Xw - Y\|^2 + \sigma_\eta^{-2} \|w\|^2\right]\right)$

Ridge loss function is (linear transform of) log-posterior with $\lambda = \sigma_\varepsilon^2 / \sigma_\eta^2$

Other penalties, e.g. Lasso (L1 regularization): $\lambda = \sum_j |w_j|$

Gaussian process regression

Rewrite ridge estimate as $\widehat{w} = X^{\mathrm{T}}(XX^{\mathrm{T}} + \lambda I_n)^{-1}Y$

$n \times n$ matrix instead of $m \times m$; more efficient if fewer cases than cues

Prediction for new stimulus: $\hat{y} = xX^{\mathrm{T}}(XX^{\mathrm{T}} + \lambda I_n)^{-1}Y$

Depends only on inner products between stimuli

$\hat{y} = K(x_{\mathrm{test}}, x_{\mathrm{train}})(K(x_{\mathrm{train}}, x_{\mathrm{train}}) + \lambda I_n)^{-1} y_{\mathrm{train}}$

Bayesian interpretation

Random function $f: \mathcal{X} \to \mathbb{R}$

Jointly Gaussian for any finite set of stimuli: $\big(f(x^1), \ldots, f(x^n)\big)^{\mathrm{T}} \sim \mathcal{N}(0, K)$ where $K_{ij} = k\big(x^i, x^j\big)$

Noisy observations: $y \sim \mathcal{N}(f(x), \lambda)$

Same generative model as ridge regression with $\sigma_\eta^2 = 1$: $(\mathbf{x}^1 w, \ldots, \mathbf{x}^n w)^{\mathrm{T}} = Xw \sim \mathcal{N}(0, XX^{\mathrm{T}})$

Posterior prediction

$$p(f_{\mathrm{all}} | X_{\mathrm{train}}, Y_{\mathrm{train}}) \propto \exp\left(-\frac{1}{2}\left(f_{\mathrm{all}}^{\mathrm{T}} K_{\mathrm{all}}^{-1} f_{\mathrm{all}} + \lambda^{-1}(f_{\mathrm{train}} - y_{\mathrm{train}})^{\mathrm{T}}(f_{\mathrm{train}} - y_{\mathrm{train}})\right)\right)$$

$$\hat{f}_{\mathrm{all}} = \left(K_{\mathrm{all}}^{-1} + \lambda^{-1}\begin{bmatrix} I_{n_{\mathrm{train}}} & 0 \\ 0 & 0 \end{bmatrix}\right)^{-1} \lambda^{-1}\begin{bmatrix} I_{n_{\mathrm{train}}} & 0 \\ 0 & 0 \end{bmatrix} y_{\mathrm{all}}$$

$$= K_{\mathrm{all,train}}\left(K_{\mathrm{train,train}} + \lambda I_{n_{\mathrm{train}}}\right)^{-1} y_{\mathrm{train}}$$