Learning with multi-modal data

Canonical correlation analysis

Geelon So, gso@seekr.com
SeekAIR — September 1, 2022

Problem

Aligning multiple views of data: given two views of data (X_1, X_2), learn separate but coordinated representations $(\tilde{X}_1, \tilde{X}_2)$ of the data with maximal linear correlation.

Canonical correlation analysis

Canonical correlation analysis (CCA, Hotelling (1936))

Given: let $(X_1, X_2) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ be a pair of random variables with covariance Σ.

Problem: find linear projections $w_i \in \mathbb{R}^{d_i}$ for $i = 1, 2$ maximizing the correlation:

$$\begin{aligned}(w_1^*, w_2^*) &:= \arg\max_{w_1, w_2} \operatorname{corr}(w_1^\top X_1, w_2^\top X_2) \\ &= \arg\max_{w_1, w_2} \frac{w_1^\top \Sigma_{12} w_2}{\sqrt{w_1^\top \Sigma_{11} w_1 \, w_2^\top \Sigma_{22} w_2}}\end{aligned}$$

Canonical correlation analysis (CCA)

Solution:

- Without loss of generality, we may assume Σ_{11} and Σ_{22} are identity matrices.
  - Otherwise, we just need to whiten the data:

$$X_1 \mapsto \Sigma_{11}^{-1/2} X_1 \qquad \text{and} \qquad X_2 \mapsto \Sigma_{22}^{-1/2} X_2.$$

- We can rewrite the CCA problem:

$$(w_1^*, w_2^*) = \underset{\|w_1\|^2 = \|w_2\|^2 = 1}{\arg\max} \; w_1^\top \Sigma_{12} w_2.$$

- This is now a familiar problem with familiar solution:
  - w_1 and w_2 are the left and right singular vectors for the top singular value of Σ_{12}.

CCA with k dimensions

Generalizing, we aim to find projections $A_i \in \mathbb{R}^{d_i \times k}$ into k dimensions:

$$(A_1^*, A_2^*) := \underset{A_1^\top \Sigma_{11} A_1 = A_2^\top \Sigma_{22} A_2 = I}{\arg\max} \operatorname{tr}\left(A_1^\top \Sigma_{12} A_2\right).$$

- This again is solved by applying SVD to $\Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1/2}$.
- Aside: this problem and solution is amenable to a kernelized approach (KCCA).

CCA Projection

Definition

We say that $\{U_1^{(j)}\}$ *and* $\{U_2^{(j')}\}$ *are* ***canonical coordinate systems*** *for* X_1 *and* X_2 *if they are a pair of orthonormal bases and satisfy:*

$$\mathrm{corr}\left(U_1^{(j)} X_1, U_2^{(j')} X_2\right) = \begin{cases} \lambda_j & j = j' \\ 0 & j \neq j'. \end{cases}$$

Without loss of generality, assume $1 \geq \lambda_1 \geq \lambda_2 \geq \cdots \geq 0$.

- A k-dimensional CCA projects onto first k basis vectors of $\{U_i^{(j)}\}$.

Theory of CCA for regression

CCA for linear regression

Linear regression in the CCA subspace:

1. Given unlabeled data $\{X^{(j)} \equiv (X_1^{(j)}, X_2^{(j)})\}$, learn CCA projections $\Pi \equiv (\Pi_1, \Pi_2)$.
2. Given (possibly distinct) labeled data $\{(X^{(j)}, Y^{(j)})\}$ perform least squares on:

$$\{(\Pi X^{(j)}, Y^{(j)})\}.$$

Theoretical result

Theorem (Informal, Foster et al. (2008))

Under either (i) redundancy or (ii) conditional independence assumptions, dimensionality reduction via CCA does not lose predictive power for linear regression.

- By performing a lower dimensional linear regression, the gain is in the reduction of sample complexity.

R^2, coefficient of determination

Recall that the **coefficient of determination** for the linear regression problem (X, Y) is the maximal correlation achievable by a linear estimator:

$$R^2_{X;Y} = \max_{\beta} \operatorname{corr}(\beta X, Y).$$

It is also equal to the fraction of explained variation in Y,

$$R^2_{X;Y} = \max_{\beta} \left(1 - \frac{\operatorname{loss}(\beta)}{\operatorname{var}(Y)}\right),$$

where $\operatorname{loss}(\beta)$ is the sum of the squared residuals $\|\beta X - Y\|^2$.

Redundancy assumption

Assumption (ε-redundancy)

Assume that the best linear predictor from each view is roughly as good as the best linear predictor from the joint views. More precisely,

$$R^2_{X_i;Y} \geq R^2_{X;Y} - \varepsilon, \qquad i = 1, 2.$$

CCA projection Π_λ

Recall that CCA finds canonical coordinate systems $\{U_1^{(j)}\}$ and $\{U_2^{(j)}\}$ so that:

$$\mathrm{corr}\left(U_1^{(j)} X_1, U_2^{(j')} X_2\right) = \begin{cases} \lambda_j & j = j' \\ 0 & j \neq j', \end{cases}$$

and $1 \geq \lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. We can define a projection onto directions that achieve a minimum threshold correlation:

$$\Pi_\lambda \text{ projects } X \text{ onto subspaces spanned by } U_i^{(j)} \text{ if } \lambda_j > \lambda.$$

Theoretical result: with redundancy assumption

Theorem

Suppose the ε-redundancy assumption holds for (X_1, X_2, Y) where $Y \in \mathbb{R}$. For all $\lambda \in [0, 1]$,

$$R^2_{\Pi_\lambda X_i; Y} \geq R^2_{X_i; Y} - \frac{4\varepsilon}{1 - \lambda}.$$

Proof sketch

By a change of basis, assume that X_1, X_2, Y are isotropic (i.e. identity covariances). Let's consider X_1 (the case for X_2 is analogous).

1. Let β_{CCA} and β_1 be the best linear predictors for $\Pi_\lambda X_1$ and X_1, respectively.
2. This means that β_{CCA} is simply the projection of β_1 onto the CCA subspace:

$$\beta_{\text{CCA}} = \beta_1 \Pi_\lambda.$$

3. Note that in the canonical coordinate system, we can write:

$$\|\beta_1 - \beta_{\text{CCA}}\|^2 = \sum_{j:\lambda_j<\lambda} \left([\beta_1]_j\right)^2,$$

where $[\beta_1]_j = \beta_1 U_1^{(j)}$ is the jth coordinate in the $\{U_1^{(j)}\}$ basis.

Proof sketch (cont.)

4. Since X_1 and Y are isotropic (e.g. $\mathrm{var}(Y) = 1$), this is precisely the amount of unexplained variation not captured by β_{CCA} when compared to β_1:

$$R^2_{X_1;Y} - R^2_{\Pi_\lambda X_1;Y} = \mathrm{loss}(\beta_{\mathrm{CCA}}) - \mathrm{loss}(\beta_1) = \|\beta_1 - \beta_{\mathrm{CCA}}\|^2.$$

5. Claim: $\|\beta_1 - \beta_{\mathrm{CCA}}\|^2 < \frac{4\varepsilon}{1-\lambda}$.
   - By ε-redundancy, both X_1 and X_2 are almost predictive of Y as $X \equiv (X_1, X_2)$. Thus:

$$\mathbb{E}\left[\left(\beta_1 X_1 - \beta_2 X_2\right)^2\right] \leq 4\varepsilon.$$

   - This implies that $[\beta_1]_j$ cannot be very large if $\lambda_j < \lambda$. Otherwise, β_1 would be much more predictive than β_2. Analytically, we get a bound:

$$\sum_j (1 - \lambda_j)\left([\beta_1]_j\right)^2 \leq 4\varepsilon.$$

□

Conditional (non)-correlation condition

Assumption

We say that H is a hidden state for (X_1, X_2, Y) if conditional on H, the triple is uncorrelated. Assume that there is a linear hidden state H such that both X_1 and X_2 are non-trivially predictive of H. Formally, for all directions w,

$$R^2_{X_i;wH} > 0$$

Theoretical result: with conditional non-correlation assumption

Theorem

Assume that there is a k-dimensional linear hidden state H for (X_1, X_2, Y). Then $\Pi_{\mathrm{CCA}} \equiv \Pi_0$ is precisely a linear projection onto a k-dimensional subspace, and:

(i) *the best linear predictor of Y with X_i is equal to the best linear predictor with $\Pi_{\mathrm{CCA}} X_i$,*

(ii) *the best linear predictor for Y with X is equal to the best linear predictor with $\Pi_{\mathrm{CCA}}(X_1, X_2)$.*

Deep canonical correlation analysis

Deep CCA (Andrew et al., 2013)

Deep canonical correlation analysis (DCCA): learn deep representations

$$f_i(\,\cdot\,;\theta_i) : \mathbb{R}^{d_i} \to \mathbb{R}^{d_o}, \qquad i = 1, 2,$$

so that the representation of the two views of data have maximal correlation. Thus:

$$(\theta_1^*, \theta_2^*) := \arg\max_{\theta_1,\theta_2} \operatorname{corr}\big(f_1(X_1), f_2(X_2)\big).$$

Training for DCCA

Andrew et al. (2013) consider fully-connected MLPs $f_i(\,\cdot\,;\theta_i)$ with hidden layers:

$$h_i^{(j)} = \sigma\big(W_i^{(j)} h_i^{(j-1)} + b_i^{(j)}\big).$$

- Each hidden layer has the same dimension as the input, $h_i^{(j)} \in \mathbb{R}^{d_i}$.
- The parameters $\theta_i = (W_i^{(j)}, b_i^{(j)})$ are initialized with a denoising autoencoder:
  - Given input data $X \in \mathbb{R}^{n\times d}$, generate noisy data $\tilde{X} \in \mathbb{R}^{n\times d}$ by adding Gaussian noise.
  - Solve for parameters W, b by optimizing:

    $$\min \|\hat{X} - X\|_F^2 + \lambda\big(\|W\|_F^2 + \|b\|^2\big),$$

    where $\hat{X} = \sigma(W\tilde{X} + b)$.
  - Sequentially generate $W^{(j)}, b^{(j)}$, where the next layer builds on the previous layer.
- Train using gradient-based optimization; however, the objective is not a sum over individual training data, so it is not amenable to mini-batches.

Experiment: MNIST digits

Views: X_1 is the left-half and X_2 is the right-half of an MNIST image.

corr	CCA	KCCA	DCCA
dev	28.1	33.5	**39.4**
test	28.0	33.0	**39.7**

Table 1: Total correlation $\text{tr}(\Sigma_{12}\Sigma_{21})$ of learned representation. Here, KCCA uses an RBF kernel.

Experiment: articulatory speech data

Views: X_1 is articulatory data (positional tracking of pellets attached to a speaker's mouth, lips, tongue, and jaw; X_2 is auditory data (MFCC encoding) of sound.

corr	CCA	KCCA (rbf)	KCCA (poly)	DCCA
fold 1	16.8	29.2	32.3	**39.2**
fold 2	15.8	25.3	29.1	**34.1**
fold 3	16.9	30.8	34.0	**39.4**
fold 4	16.6	28.6	32.4	**37.1**
fold 5	16.2	26.2	29.9	**34.0**

Table 2: Total correlation $\mathrm{tr}(\Sigma_{12}\Sigma_{21})$ of learned representation; 5 independent folds of data.

Applications

Multi-model emotional recognition (Liu et al., 2019)

Problem. Emotional recognition from multi-modal datasets:

- SEED: EEG + eye movement
- DEAP: EEG + peripheral physiological signals (EOG, EMG, GSR, respiration belt, and plethysmograph)
- DREAMER: two-channel ECG

Multi-model emotional recognition (Liu et al., 2019)

Approach.

1. Apply DCCA to learn representations for both signals
2. Construct joint representation via a convex combination of the learned representation
3. Apply SVM for downstream classification task

Experimental results: accuracy

Method	Accuracy	Std.
Concatenation	83.70	–
MAX	81.71	–
FuzzyIntegral	87.59	19.87
BDAE	91.01	8.91
DGCNN	90.40	8.49
Bimodal-LSTM	93.97	7.03
DCCA	**93.58**	**6.16**

Table 3: Comparison of methods on SEED.

Experimental results: noise robustness

RECOGNITION RESULTS (MEAN/STD (%)) AFTER REPLACING DIFFERENT PROPORTIONS OF EEG FEATURES WITH VARIOUS TYPES OF NOISE. FIVE FUSION STRATEGIES UNDER VARIOUS SETTINGS ARE COMPARED, AND THE BEST RESULTS FOR EACH SETTING ARE IN BOLD

Methods	No noise	Gaussian			Gamma			Uniform		
		10%	30%	50%	10%	30%	50%	10%	30%	50%
Concatenation	73.65/8.90	70.08/8.79	63.13/9.05	58.32/7.51	69.71/8.51	62.93/8.46	57.97/8.14	71.24/10.56	66.46/9.38	61.82/8.35
MAX	73.17/9.27	67.67/8.38	58.29/8.41	51.08/7.00	67.24/10.27	59.18/9.77	50.56/6.82	67.51/9.72	60.14/9.28	52.71/7.84
FuzzyIntegral	73.24/8.72	69.42/8.92	62.98/7.52	57.69/8.70	69.35/8.70	62.64/8.90	57.56/7.19	69.16/8.16	64.86/9.37	60.47/8.32
BDAE	79.70/**4.76**	47.82/7.77	45.89/7.82	44.51/7.43	45.27/**6.68**	45.75/7.91	45.09/8.37	46.13/8.17	46.88/7.14	45.50/9.59
DCCA-0.3	79.04/7.32	76.57/7.63	**73.00**/7.36	**69.56**/7.02	76.87/7.99	**73.06**/7.00	**70.03**/7.17	75.69/**6.34**	**73.22**/6.50	**70.01**/6.66
DCCA-0.5	81.62/6.95	**77.92/6.63**	71.77/6.55	65.21/6.24	**78.29**/7.38	72.45/6.14	65.75/6.08	**78.28**/7.16	73.20/6.96	68.01/7.08
DCCA-0.7	**83.08**/7.11	76.27/7.02	68.48/**5.54**	57.63/**5.15**	76.82/7.01	68.54/**6.02**	58.58/**5.44**	77.39/8.43	69.80/**5.63**	61.58/**5.38**

Figure 1: They compare noise robustness of DCCA with that of existing methods when adding various amounts of noise to the SEED-V dataset.

Related work

Related work: co-training

Blum and Mitchell (1998) introduces the co-training framework:

- supervised learning problem: learn (X, Y) with two views of $X \equiv (X_1, X_2)$
- there is a joint concept class $f \equiv (f_1, f_2) \in C_1 \times C_2$
- realizability assumption: there is some f^* such that:

  $$f_1(X_1) = f_2(X_2) = Y$$

  for all instances (X, Y) generated by nature
- they prove PAC-style results under the redundancy assumption:
  - it is possible to learn Y by observing only X_1 or only X_2

Related work: multi-view redundancy for contrastive learning

Tosh et al. (2021) show similar results for the contrastive learning setting:

- supervised learning problem: learn (X, Y) with two views of $X \equiv (X_1, X_2)$
- first solve the contrastive estimation problem:
  - construct supervised learning task from unlabeled data of the form

    $$(X_1, X_2, +1) \qquad \text{and} \qquad (X_1, \tilde{X}_2, -1),$$

    where (X_1, X_2) and $(\tilde{X}_1, \tilde{X}_2)$ are i.i.d. draws from $\mathcal{X}$
- sample a set of landmarks $X_2^{(1)}, \ldots, X_2^{(m)}$ to generate *landmark embeddings* $g^*(x_1)$

  $$g^*(x_1)_j = \frac{p(x_1 | X_2^{(j)})}{p(x_1)}$$

- they prove that if a redundancy assumption is satisfied, the linear functions on landmark embeddings can perform well

Questions

- If views are produced by randomly masking subsets of features, is there a relationship to dropout?
- When are situations where DCCA would not be helpful?
- Could we learn representations where many tasks correspond to a different linear view of the data?
- Does introducing noise contrastive estimation help? Minimize correlation across unrelated views?

References

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

Dean P Foster, Sham M Kakade, and Tong Zhang. Multi-view dimensionality reduction via canonical correlation analysis. *Technical Report TR-2009-5, TTI-Chicago*, 2008.

Harold Hotelling. Relations between two sets of variates. In *Biometrika*. 1936.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Multimodal emotion recognition using deep canonical correlation analysis. *arXiv preprint arXiv:1908.05349*, 2019.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206. PMLR, 2021.