Nonlinear dimensionality reduction NeRV method

Lauri Oksanen

November 20, 2007

Outline

Measure for Goodness of Visualization

NeRV Method

Reference: Venna, J., Kaski, S. (2007). Nonlinear Dimensionality Reduction as Information Retrieval.

Starting-point

- We want to map data $x_1, \ldots, x_N \in X$ into a lower dimensional space $x_i \mapsto y_i \in Y$ for visualization.
- Distributions p^i model the *neighborhood relations in X* i.e.

$$p^i_j := P(x_j \text{ is the nearest neighbor of } x_i), \quad j \neq i.$$

- Distributions q^i model how the *neighborhood relations* are perceived *in Y* i.e.

$$q^i_j := P(y_j \text{ looks like the nearest neighbor of } y_i), \quad j \neq i.$$

- Mapping $x_i \mapsto y_i$ is *optimal* if $p^i = q^i$ for all $i = 1, \ldots, N$.

Measuring the difference between p^i and q^i

- Kullback-Leibler (KL) divergence $D(p,q)$ is a standard information theoretic measure of the difference between two distributions p, q.
- Assuming that $q_j > 0$ whenever $p_j > 0$, KL divergence can be defined
$$D(p,q) := \sum_j p_j \log \frac{p_j}{q_j}.$$
- KL divergence is not symmetric, i.e. it can be that $D(p,q) \neq D(q,p)$.
- Should we use $D(p^i, q^i)$, $D(q^i, p^i)$ or both?

Interpretation of $D(p^i, q^i)$

- Suppose that we are using simple model

$$p_j^i := \begin{cases} 1/k, & x_j \in N_k(x_i) \\ 0, & \text{otherwise}, \end{cases}$$

$$q_j^i := \begin{cases} a \approx 1/r, & y_j \in N_r(y_i) \\ b \approx 0, & \text{otherwise}, \end{cases}$$

where $N_k(x_i)$ is the set of k nearest neighbors of x_i.

- Now

$$D(p^i, q^i) = \sum_{j: y_j \in N_r(y_i), x_j \in N_k(x_i)} \frac{1}{k} \log \frac{1}{ka} + \sum_{j: y_j \notin N_r(y_i), x_j \in N_k(x_i)} \frac{1}{k} \log \frac{1}{kb}.$$

Interpretation of $D(p^i, q^i)$ (continues)

- Now
$$D(p^i, q^i) = \frac{1}{k}(\log \frac{1}{ka} N_{TP} + \log \frac{1}{kb} N_{MISS}),$$
where N_{TP} is the number of points in $N_k(x_i)$ that are mapped into $N_r(y_i)$ (true positives) and N_{MISS} is the number of points in $N_k(x_i)$ that are not mapped into $N_r(y_i)$ (misses).
- Let $b \to 0$. Then $a \to 1/r$ and
$$D(p^i, q^i) \to \frac{1}{k}(\log \frac{r}{k} N_{TP} + \infty N_{MISS}).$$
Hence $D(p^i, q^i) \propto \frac{N_{MISS}}{k}$ when b is small.

Interpretation of $D(q^i, p^i)$

- Suppose that $p^i = a1_{N_k(x_i)} + b1_{N_k(x_i)^c}$ and $q^i = \frac{1}{r}1_{N_r(y_i)}$.
- Now
$$D(q^i, p^i) = \frac{1}{r}(\log \frac{1}{ra} N_{TP} + \log \frac{1}{rb} N_{FP}),$$
where N_{TP} is the number of true positives as before and N_{FP} is the number of points not in $N_k(x_i)$ that are mapped into $N_r(y_i)$ (false positives).
- Let $b \to 0$. Then $a \to 1/k$ and
$$D(q^i, p^i) \to \frac{1}{r}(\log \frac{k}{r} N_{TP} + \infty N_{FP}).$$
Hence $D(q^i, p^i) \propto \frac{N_{FP}}{r}$ when b is small.

Summing up $D(p^i, q^i)$ and $D(q^i, p^i)$

- For uniform distributions concentrated on $N_k(x_i)$ and $N_r(y_i)$ we have approximately that
  - $D(p^i, q^i)$ is proportional to the frequency of misses
  - $D(q^i, p^i)$ is proportional to the frequency of false positives
- It makes sense to penalize for the both
- Define a measure for the difference of p^i and q^i

$$\lambda D(p^i, q^i) + (1 - \lambda) D(q^i, p^i),$$

where $\lambda \in [0, 1]$.

Definition of NeRV Method

- Define the model

$$p^i_j := C_i \exp(-d(x_i, x_j)^2/\sigma_i^2),$$
$$q^i_j := C_i \exp(-d(y_i, y_j)^2/\sigma_i^2).$$

- Initialization
  - Select parameters $\lambda \in [0, 1]$ and σ_i,
  - Select initial values $y_1, \ldots, y_N$.
- Minimize the cost function

$$E(y_1, \ldots, y_N) := \sum_i [\lambda D(p^i, q^i) + (1 - \lambda) D(q^i, p^i)]$$

using conjugate gradient (CG) method.

Properties of NeRV

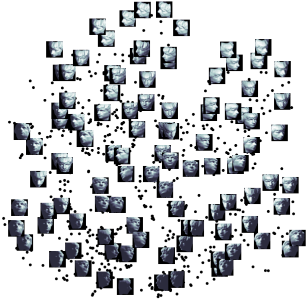
+ Method has theoretical justification as we have seen.
+ Parameter λ controls the trade-off between misses and false positives.
  - This is because the cost function E can be interpreted as a smoothed version of the uniform case.
+ Effective heuristic to avoid local minima in optimization exists.
  - Start with large width of Gaussian neighborhoods σ_i^2 and decrease it after each CG step.
  - When the final value of σ_i^2 is reached continue with normal CG.

\- Conjugate gradient step is of complexity $\mathcal{O}(N^3)$.

Example: Projecting 3D shere into 2D

- Original 3D coordinates govern the rotation, scale and elongation of the markers.
- On the left $\lambda = 0$, false positives are avoided, sphere is splitted open.
- On the right $\lambda = 1$, misses are avoided, sphere is compressed flat.

Example: Projecting faces into 2D



- Faces form a 3D manifold (pose up-down, pose left-right, light left-right) in the 4094 (64x64) dimensional image space.
- This manifold is projected into 2D space using different methods.

Example: Projecting faces into 2D (continues)

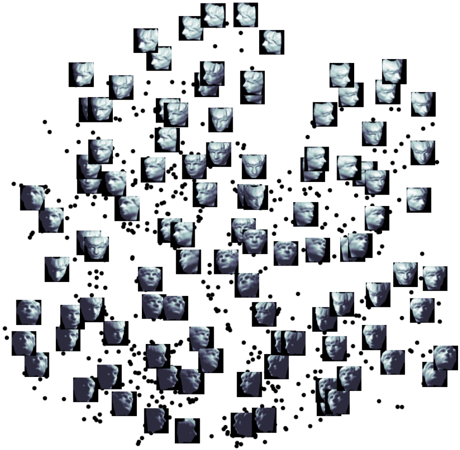


Figure: Projection using NeRV, $\lambda = 0.1$.

Example: Projecting faces into 2D (continues)

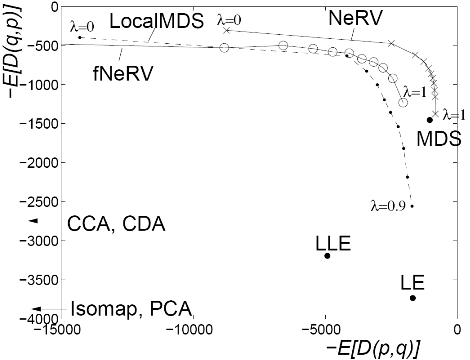


Figure: Estimated KL divergences using 20 nearest neighbors and the image space as the input space.

Example: Projecting faces into 2D (continues)

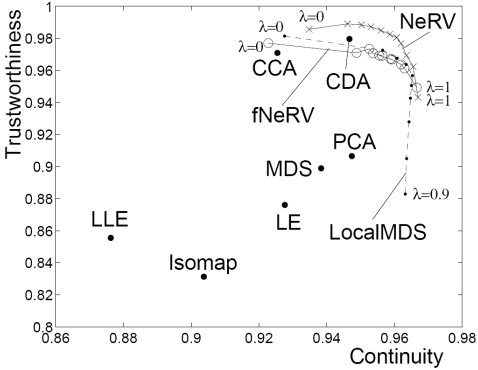


Figure: Trustworthiness-continuity using 20 nearest neighbors and the image space as the input space.

Example: Projecting faces into 2D (continues)

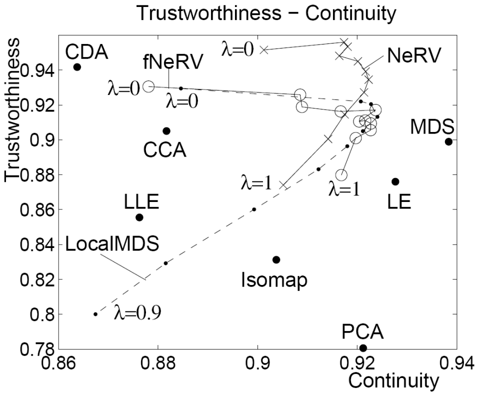


Figure: Trustworthiness-continuity using 20 nearest neighbors and the known pose/lighting space as the input space.