t_sne

Luis Angel Rodriguez Garcia

3/5/2022

Introduction

We are going to play with the Iris dataset:

```
X <- iris %>% dplyr::select(-Species) %>% as.matrix()
n <- nrow(X)
p <- ncol(X)
```

Euclidean distance

The way to calculate the Euclidean distance:

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 = \|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2 - 2\mathbf{x}_i'\mathbf{x}_j$$

where $\|\mathbf{x}_i\| = \sqrt{x_{i,1}^2 + \cdots + x_{i,p}^2}$ and $\mathbf{x}_k^2 = \mathbf{x}_k'\mathbf{x}_k = x_{k1}^2 + ... + x_{kp}^2$.

The following method apply the formula above:

```
x_diff <- function(X) {
  n <- nrow(X)
  sum_x <- apply(X^2, MARGIN=1, FUN=sum)
  sum_x_m <- t(matrix(replicate(n, sum_x), byrow=T, nrow=n))
  cross_times_minus_2 <- -2 * (X %*% t(X))
  D <- t(cross_times_minus_2 + sum_x_m) + sum_x_m
  D <- round(D, digits=4)
}
```

Perplexity

$$Perp_i = 2^{H_i}$$

Where H_i is the Shannon entropy in the point x_i of the conditional probability:

$$
\begin{aligned}
H_i &= -\sum_{j\neq i} p_{j|i}\log(p_{j|i}) \\
&= -\sum_{j\neq i} p_{j|i}\log(p_{j,i}/p_i) \\
&= -\sum_{j\neq i} p_{j|i}(\log(p_{j,i}) - \log(p_i)) \\
&= -\sum_{j\neq i} p_{j|i}(\log(e^{-||x_i-x_j||^2/2\sigma^2}) - \log(\sum_{k\neq i} e^{-||x_i-x_j||^2/2\sigma^2})) \\
&= -\sum_{j\neq i} p_{j|i}((-||x_i-x_j||^2/2\sigma^2) - \log(\sum_{k\neq i} e^{-||x_i-x_j||^2/2\sigma^2})) \\
&= \sum_{j\neq i} p_{j|i}(\log(S_i) + ||x_i-x_j||^2\frac{1}{2\sigma^2}) \\
&= \log(S_i)\sum_{j\neq i} p_{j|i} + \frac{1}{2\sigma^2}\sum_{j\neq i} p_{j|i}||x_i-x_j||^2) \\
&= \log(S_i) + \frac{1}{2\sigma^2}\sum_{j\neq i} p_{j|i}||x_i-x_j||^2
\end{aligned}
\tag{1}
$$

Where $S_i = \sum_{k\neq i} e^{-||x_i-x_j||^2/2\sigma^2}$ and $\sum_{j\neq i} p_{j|i} = 1$

In order to proceed with the optimization of the variance, we are going to define this term $\frac{1}{2\sigma^2}$ as the parameter β.

```
entropy_beta <- function(D_i, beta=1) {
  P_i <- exp(-D_i * beta)
  sum_p_i <- sum(P_i)
  H_i <- log(sum_p_i) + (beta * sum(D_i * P_i) /sum_p_i)
  P_i <- P_i / sum_p_i
  return(list(entropy=H_i, probs=P_i))
}
```

The goal is to adjust the variability so that the perplexity at each point is the same. The perplexity is a way to measure the effective number of neighbors of a point. We are going to perform a binary search to get the probabilities in such a way that the conditional Gaussian has the same perplexity.

```
index_except_i <- function(i, n) {
  index <- c(seq(1,i-1),seq(i+1,n))
  if (i == 1) {
    index <- 2:n
  } else if (i == n) {
    index <- 1:(n-1)
  }
  return(index)
}

binary_search <- function(h_diff, beta, i, beta_min, beta_max) {
  if(h_diff > 0) {
    beta_min = beta[i]
    if(beta_max == -Inf || beta_max == Inf) {
      beta[i] <- beta[i] * 2
    } else {
```

```r
      beta[i] <- (beta[i] + beta_max) / 2
    }
  } else {
    beta_max = beta[i]
    if(beta_min == -Inf || beta_min == Inf) {
      beta[i] <- beta[i] / 2
    } else {
      beta[i] <- (beta[i] + beta_min) / 2
    }
  }
  return(list(beta=beta, min=beta_min, max=beta_max))
}

binary_search_optimization <- function(D_i, i, beta, h_star, prob_star, log_perp,
                                       tolerance=1e-5) {
  beta_min <- -Inf
  beta_max <- Inf
  tries <- 0
  h_diff <- h_star - log_perp

  while(abs(h_diff) > tolerance && tries < 50) {
    beta_opt <- binary_search(h_diff, beta, i, beta_min, beta_max)
    beta <- beta_opt$beta; beta_min <- beta_opt$min; beta_max <- beta_opt$max

    res_loop <- entropy_beta(D_i, beta[i])
    h_star <- res_loop$entropy; prob_star <- res_loop$probs

    h_diff <- h_star - log_perp
    tries <- tries + 1
  }
  return(list(probs=prob_star, beta=beta))
}
```

Once we have defined these two methods, we are able to obtain the high dimensional properties:

```r
high_dimension_probs <- function(X=matrix(), tolerance=1e-5, perplexity=30) {
  n <- nrow(X)
  p <- ncol(X)

  D <- x_diff(X)

  P <- matrix(0, nrow=150, ncol=150)
  beta <- rep(1, n)
  log_perp <- log(perplexity)

  for(i in seq_len(n)) {
    column_index <- index_except_i(i, n)
    D_i <- D[i, column_index]

    res <- entropy_beta(D_i, beta[i])
    h_star <- res$entropy
    prob_star <- res$probs

    h_diff <- h_star - log_perp
```

```
    res_opt = binary_search_optimization(D_i, i, beta, h_star, prob_star,
                                         log_perp, tolerance)
    prob_star = res_opt$probs; beta = res_opt$beta

    P[i, column_index] <- prob_star
  }
  print("Values of sigma for each x_i")
  print(sqrt(1/beta))
  print(sprintf("Mean value of sigma: %01.2f", mean(sqrt(1/beta))))
  return(P)
}
```

The version of the t-SNE is the symmetric one, that has the property that $p_{ij} = p_{ji}$ and $q_{ij} = q_{ji} \quad \forall i, j$. Therefore, we define $p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n}$.

```
symmetric_probs <- function(P) {
  P = (P + t(P)) / (2*nrow(P))
  return(P)
}
```

In order to initialize the lower dimension probability matrix, we are going to use the method `mvtnorm::rmvnorm` as it is described in the paper: $\mathcal{Y}^0 = \{y_1, ..., y_n\} \sim \mathcal{N}(0, 10^{-4}\mathbf{I}_n)$ which is assigned to $\mathcal{Y}^1$ and $\mathcal{Y}^2$ (the first two initial states).

Gradient Descent

$$
\begin{aligned}
C &= KL(\mathbf{P} \| \mathbf{Q}) \\
&= \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \\
&= \sum_i \sum_j p_{ij} \left(\log p_{ij} - \log q_{ij}\right) \\
&= \sum_i \sum_j p_{ij} \log p_{ij} - p_{ij} \log q_{ij}
\end{aligned}
\tag{2}
$$

We define these two auxiliary variables $d_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\|$ and $Z = \sum_{k \neq l} \left(1 + d_{kl}^2\right)^{-1}\big)$

$$
\begin{aligned}
\frac{\partial C}{\partial \mathbf{y}_i} &= \sum_{j \neq i} \left[\frac{\partial C}{\partial d_{ij}} \frac{\partial d_{ij}}{\partial \mathbf{y}_i} + \frac{\partial C}{\partial d_{ji}} \frac{\partial d_{ji}}{\partial \mathbf{y}_i}\right] \\
&= \sum_{j \neq i} \left[\frac{\partial d_{ij}}{\partial \mathbf{y}_i} \left(\frac{\partial C}{\partial d_{ij}} + \frac{\partial C}{\partial d_{ji}}\right)\right]
\end{aligned}
\tag{3}
$$

Recall that $\frac{\partial}{\partial x} \sqrt{g(x)} = \frac{1}{2\sqrt{g(x)}} g'(x)$, $\frac{\partial}{\partial \mathbf{y}} \|\mathbf{y}\|^2 = 2\mathbf{y}$ and $d_{ij} = d_{ji}$

$$
\begin{aligned}
\frac{\partial d_{ij}}{\partial \mathbf{y}_i} &= \frac{\partial}{\partial y_i}\|\mathbf{y}_i - \mathbf{y}_j\| \\
&= \frac{\partial}{\partial \mathbf{y}_i}(\|\mathbf{y}_i\|^2 + \|\mathbf{y}_j\|^2 - 2\mathbf{y}_i'\mathbf{y}_j)^{\frac{1}{2}} \\
&= \frac{1}{2}\frac{1}{d_{ij}}\frac{\partial}{\partial \mathbf{y}_i}(\|\mathbf{y}_i\|^2 + \cancel{\|\mathbf{y}_j\|^2} - 2\mathbf{y}_i'\mathbf{y}_j) \\
&= \frac{1}{2}\frac{1}{d_{ij}}(2\mathbf{y}_i - 2\mathbf{y}_j) \\
&= \frac{(\mathbf{y}_i - \mathbf{y}_j)}{d_{ij}} \\
&= \frac{\partial d_{ji}}{\partial \mathbf{y}_i}
\end{aligned}
\tag{4}
$$

```r
dij.1 <- function(i, j) {
  sqrt(norm(i, type="2")^2 + norm(j, type="2")^2 - 2 * (t(i) %*% j))
}

dij.2 <- function(i, j) {
  norm(i-j, type="2")
}

y <- data.matrix(iris)[, -5]

# For rows 2 and 3
result1 <- as.numeric(round(jacobian(func=dij.2, x=y[2,], j=y[3,]), digits=7))
result2 <- as.numeric((y[2,]-y[3,])/norm(y[2,]-y[3,], type="2"))
all.equal(result2, result1) # checked
```

```
## [1] "Mean relative difference: 6e-08"
```

```r
# For row 3 and 2
result3 <- as.numeric(round(jacobian(func=dij.2, x=y[3,], i=y[2,]), digits=7))
result4 <- as.numeric((y[3,]-y[2,])/norm(y[2,]-y[3,], type="2"))
all.equal(result4, result3) # checked
```

```
## [1] "Mean relative difference: 6e-08"
```

```r
# Checked that both d(d_ij)/dy_i = d(d_ji)/dy_i
```

Recall that $d_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\|$ and $Z = \sum_{k \neq l}\left(1 + d_{kl}^2\right)^{-1}\big)$:

$$
\begin{aligned}
\frac{\partial C}{\partial d_{ij}} &= \frac{\partial}{\partial d_{ij}} \sum_{k \neq l} \cancel{p_{kl} \log p_{kl}} - p_{kl} \log q_{kl} \\
&= \frac{\partial}{\partial d_{ij}} \sum_{k \neq l} -p_{kl} \log q_{kl} = -\sum_{k \neq l} p_{kl} \frac{\partial (\log q_{kl})}{\partial d_{ij}} \\
&= -\sum_{k \neq l} p_{kl} \frac{\partial (\log \frac{q_{kl} Z}{Z})}{\partial d_{ij}} \\
&= -\sum_{k \neq l} p_{kl} \left[\frac{\partial (\log q_{kl} Z - \log Z)}{\partial d_{ij}} \right] \\
&= -\sum_{k \neq l} p_{kl} \left[\frac{\partial (\log q_{kl} Z)}{\partial d_{ij}} - \frac{\partial \log Z}{\partial d_{ij}} \right] \\
&= -\sum_{k \neq l} p_{kl} \left[\frac{1}{q_{kl} Z} \frac{\partial (q_{kl} Z)}{\partial d_{ij}} - \frac{1}{Z} \frac{\partial Z}{\partial d_{ij}} \right] \\
&= -\sum_{k \neq l} p_{kl} \left[\frac{1}{q_{kl} Z} \frac{\partial \left(\frac{(1+d_{ij}^2)^{-1}}{\cancel{\sum_{k \neq l} (1+d_{kl}^2)^{-1}}} \cancel{\sum_{k \neq l} (1+d_{kl}^2)^{-1}} \right)}{\partial d_{ij}} - \frac{1}{Z} \frac{\partial (\sum_{k \neq l} (1+d_{kl}^2)^{-1})}{\partial d_{ij}} \right] \\
&= 2 \frac{p_{ij}}{q_{ij} Z} (1+d_{ij}^2)^{-2} d_{ij} - 2 \sum_{k \neq l} p_{kl} \frac{(1+d_{ij}^2)^{-1}}{Z} (1+d_{ij}^2)^{-1} d_{ij} \\
&= 2 \frac{p_{ij}}{\frac{\cancel{(1+d_{ij}^2)^{-1}}}{\cancel{Z}} \cancel{Z}} (1+d_{ij}^2)^{-\cancel{2}1} d_{ij} - 2 \cancel{\sum_{k \neq l} p_{kl}}\, q_{ij} (1+d_{ij}^2)^{-1} d_{ij} \\
&= 2 p_{ij} (1+d_{ij}^2)^{-1} d_{ij} - 2 q_{ij} (1+d_{ij}^2)^{-1} d_{ij} \\
&= 2 (p_{ij} - q_{ij}) (1+d_{ij}^2)^{-1} d_{ij}
\end{aligned}
\tag{5}
$$

$$
\begin{aligned}
\frac{\partial C}{\partial y_i} &= \sum_{j \neq i} \left[\frac{\partial C}{\partial d_{ij}} \frac{\partial d_{ij}}{\partial y_i} + \frac{\partial C}{\partial d_{ji}} \frac{\partial d_{ji}}{\partial y_i} \right] \\
&= \sum_{j \neq i} \left[\frac{\partial d_{ij}}{\partial y_i} \left(\frac{\partial C}{\partial d_{ij}} + \frac{\partial C}{\partial d_{ij}} \right) \right] \\
&= 2 \sum_{j \neq i} \left[\frac{\partial C}{\partial d_{ij}} \right] \frac{\mathbf{y}_i - \mathbf{y}_j}{d_{ij}} \\
&= 2 \sum_{j \neq i} \left[2 (p_{ij} - q_{ij}) (1+d_{ij}^2)^{-1} d_{ij} \right] \frac{\mathbf{y}_i - \mathbf{y}_j}{d_{ij}} \\
&= 4 \sum_{j \neq i} (p_{ij} - q_{ij}) (1+d_{ij}^2)^{-1} \cancel{d_{ij}} \frac{\mathbf{y}_i - \mathbf{y}_j}{\cancel{d_{ij}}} \\
&= 4 \sum_{j \neq i} (p_{ij} - q_{ij}) (1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1} (\mathbf{y}_i - \mathbf{y}_j)
\end{aligned}
\tag{6}
$$