El enfoque de suficiencia en la reducción de dimensiones para el reconocimiento de patrones

Diego Tomassi

Trabajo conjunto con Liliana Forzani y Diego Milone

11 de junio de 2010

Organización

1. Introducción
2. Antecedentes
3. El enfoque de reducción suficiente de dimensiones
4. Simulaciones
5. Comentarios finales

Organización

1. Introducción
2. Antecedentes
3. El enfoque de reducción suficiente de dimensiones
4. Simulaciones
5. Comentarios finales

El problema de clasificación

Ejemplos

- Detección de patologías para asistencia de diagnóstico médico.
- Reconocimiento de rostros y aplicaciones biométricas en general.
- Reconocimiento de escritura manuscrita (tablet PC).
- Reconocimiento del habla.
- Filtros anti-spam.

El problema de clasificación

Sean $Y = 1, 2, \ldots, h$ las clases y $\mathbf{X} \in \mathbb{R}^p$ las características.

Un clasificador es una función $f : \mathbb{R}^p \rightarrow \{1, 2, \ldots, h\}$ tal que la predicción $\widehat{Y} = f(\mathbf{X})$ tiene una mínima probabilidad de error.

El problema de clasificación

Sean $Y = 1, 2, \ldots, h$ las clases y $\mathbf{X} \in \mathbb{R}^p$ las características.

Un clasificador es una función $f : \mathbb{R}^p \to \{1, 2, \ldots, h\}$ tal que la predicción $\widehat{Y} = f(\mathbf{X})$ tiene una mínima probabilidad de error.

Si conocemos $p(Y, \mathbf{X})$ la regla óptima es

$$f(\mathbf{X}) = \operatorname*{arg\,max}_y \; p(Y = y|\mathbf{X})$$

El problema de clasificación

- Suponemos modelos paramétricos para $p(\mathbf{X}|Y=y)$.
- Usamos un conjunto de ejemplos $\{(Y_i, \mathbf{X}_i)\}$ para estimar.

El problema de clasificación

- Suponemos modelos paramétricos para $p(\mathbf{X}|Y = y)$.
- Usamos un conjunto de ejemplos $\{(Y_i, \mathbf{X}_i)\}$ para estimar.

Definidos los modelos, el diseño del clasificador se reduce a un problema de estimación de parámetros.

El problema de clasificación

- Suponemos modelos paramétricos para $p(\mathbf{X}|Y=y)$.
- Usamos un conjunto de ejemplos $\{(Y_i, \mathbf{X}_i)\}$ para estimar.

Definidos los modelos, el diseño del clasificador se reduce a un problema de estimación de parámetros.

¿Qué modelos se usan?

- Gaussianas
- Mezclas de Gaussianas
- Modelos ocultos de Markov
- ...

El problema de clasificación

- Suponemos modelos paramétricos para $p(\mathbf{X}|Y = y)$.
- Usamos un conjunto de ejemplos $\{(Y_i, \mathbf{X}_i)\}$ para estimar.

Definidos los modelos, el diseño del clasificador se reduce a un problema de estimación de parámetros.

¿Qué modelos se usan?

- Gaussianas

$$\mathbf{X}|Y = y \sim \mathcal{N}(\mathbf{X}|\mu_y, \Delta_y)$$

Reducción de dimensiones

¿Para qué?

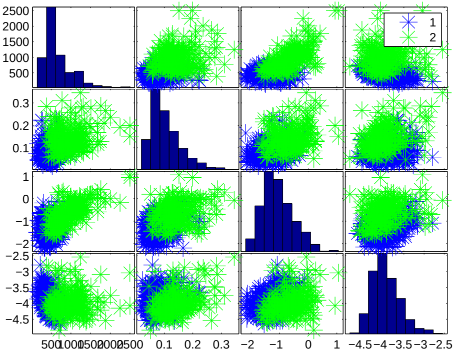
- Disminuir varianza de estimadores.

Reducción de dimensiones

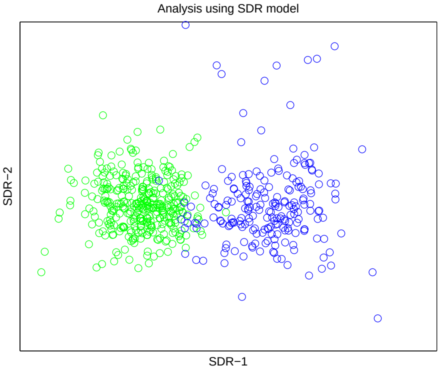
¿Para qué?

- Disminuir varianza de estimadores.
- Facilitar la decisión del clasificador.

Reducción de dimensiones



Reducción de dimensiones



Reducción de dimensiones

¿Para qué?

- Facilitar la decisión del clasificador.
- Disminuir varianza de estimadores.
- Reducir costo computacional.

Reducción de dimensiones

¿Para qué?

- Facilitar la decisión del clasificador.
- Disminuir varianza de estimadores.
- Reducir costo computacional.

¿Qué buscamos?

Una reducción $R : \mathbb{R}^p \to \mathbb{R}^d, d \leq p$ tal que para cada $\mathbf{X}$

$$p(Y = y|R(\mathbf{X})) = p(Y = y|\mathbf{X})$$

Organización

Métodos usados comúnmente

Consideremos:

- Reducciones lineales $R(\mathbf{X}) = \boldsymbol{\theta}^T \mathbf{X}$, con $\boldsymbol{\theta} \in \mathbb{R}^{p \times d}$.
- Estimación de máxima verosimilitud.

Métodos usados comúnmente

- Análisis discriminante lineal (LDA)

Métodos usados comúnmente

- Análisis discriminante lineal (LDA)
- Análisis discriminante lineal para modelos heteroscedásticos (HLDA).

Sea $\boldsymbol{\Theta} = (\boldsymbol{\theta}\ \boldsymbol{\theta}_0) \in \mathbb{R}^{p\times p}, \boldsymbol{\theta} \in \mathbb{R}^{p\times d}, \boldsymbol{\theta}^T\boldsymbol{\theta} = \mathbf{I}_d, \boldsymbol{\theta}^T\boldsymbol{\theta}_0 = 0.$
Supongamos $\boldsymbol{\Theta}^T\mathbf{X} \sim \mathcal{N}(\mu_y^*, \Delta_y^*)$, con

$$\mu_y^* = \begin{pmatrix} \boldsymbol{\theta}^T\mu_y \\ \boldsymbol{\theta}_0^T\mu \end{pmatrix} \qquad \Delta_y^* = \begin{pmatrix} \Omega_y & \mathbf{0} \\ \mathbf{0} & \Omega_0 \end{pmatrix}$$

Métodos usados comúnmente

- Análisis discriminante lineal (LDA)
- Análisis discriminante lineal para modelos heteroscedásticos (HLDA).

Sea $\boldsymbol{\Theta} = (\boldsymbol{\theta}\ \boldsymbol{\theta}_0) \in \mathbb{R}^{p\times p}, \boldsymbol{\theta} \in \mathbb{R}^{p\times d}, \boldsymbol{\theta}^T\boldsymbol{\theta} = \mathbf{I}_d, \boldsymbol{\theta}^T\boldsymbol{\theta}_0 = 0.$
Supongamos $\boldsymbol{\Theta}^T\mathbf{X} \sim \mathcal{N}(\mu_y^*, \Delta_y^*)$, con

$$\mu_y^* = \begin{pmatrix} \boldsymbol{\theta}^T\mu_y \\ \boldsymbol{\theta}_0^T\mu \end{pmatrix} \qquad \Delta_y^* = \begin{pmatrix} \Omega_y & \mathbf{0} \\ \mathbf{0} & \Omega_0 \end{pmatrix}$$

HLDA busca $\boldsymbol{\Theta}$ que maximiza

$$\mathcal{L}_{\mathsf{HLDA}}(\boldsymbol{\theta}) = -\frac{n}{2}\log|\boldsymbol{\theta}^T\tilde{\Sigma}^{-1}\boldsymbol{\theta}| - \frac{1}{2}\sum_{y=1}^{h} n_y \log|\boldsymbol{\theta}^T\tilde{\Delta}_y\boldsymbol{\theta}|.$$

Nos preguntamos...

- ¿Se satisface $p(Y = y|\boldsymbol{\theta}_{\mathsf{HLDA}}^T \mathbf{X}) = p(Y = y|\mathbf{X})$ para todo $\mathbf{X}$?

- ¿Es span$(\boldsymbol{\theta}_{\mathsf{HLDA}})$ el menor subespacio que conserva la información?

Organización

Resultados conocidos

Cook y Yin, 2001: *para* $\mathbf{X}|Y = y \sim \mathcal{N}(\mu_y, \Delta_y)$*:*

$$f(\boldsymbol{\theta}^T\mathbf{X}) = f(\mathbf{X}) \Leftrightarrow X|(Y, \boldsymbol{\theta}^T\mathbf{X}) \sim \mathbf{X}|\boldsymbol{\theta}^T\mathbf{X}$$

Resultados conocidos

Cook y Yin, 2001: *para* $\mathbf{X}|Y = y \sim \mathcal{N}(\mu_y, \Delta_y)$*:*

$$f(\boldsymbol{\theta}^T\mathbf{X}) = f(\mathbf{X}) \Leftrightarrow X|(Y, \boldsymbol{\theta}^T\mathbf{X}) \sim \mathbf{X}|\boldsymbol{\theta}^T\mathbf{X}$$

Resultados conocidos

Cook y Yin, 2001: *para* $\mathbf{X}|Y = y \sim \mathcal{N}(\mu_y, \Delta_y)$*:*

$$f(\boldsymbol{\theta}^T\mathbf{X}) = f(\mathbf{X}) \Leftrightarrow X|(Y, \boldsymbol{\theta}^T\mathbf{X}) \sim \mathbf{X}|\boldsymbol{\theta}^T\mathbf{X}$$

Cook y Forzani, 2008:

Supongamos que $\mathbf{X}|Y = y \sim \mathcal{N}(\mu_y, \Delta_y), y = 1, 2, \ldots, h$*. Sea* $\Delta = E(\Delta_y)$*,* $\boldsymbol{\theta} \in \mathbb{R}^{p\times d}$ *una matriz semiortogonal y sea* $(\boldsymbol{\theta}\ \boldsymbol{\theta}_0) \in \mathbb{R}^{p\times p}$ *una matriz ortogonal.* $\boldsymbol{\theta}^T\mathbf{X}$ *es una reducción lineal suficiente si y sólo si*

1. $\mathrm{span}(\mu_y - \mu) \subseteq \Delta\mathrm{span}(\boldsymbol{\theta})$.
2. $\boldsymbol{\theta}_0^T\Delta_y^{-1}$ es constante.

Resultados conocidos

Cook y Yin, 2001: *para* $\mathbf{X}|Y = y \sim \mathcal{N}(\mu_y, \Delta_y)$*:*

$$f(\boldsymbol{\theta}^T\mathbf{X}) = f(\mathbf{X}) \Leftrightarrow X|(Y, \boldsymbol{\theta}^T\mathbf{X}) \sim \mathbf{X}|\boldsymbol{\theta}^T\mathbf{X}$$

Cook y Forzani, 2008:

Supongamos que $\mathbf{X}|Y = y \sim \mathcal{N}(\mu_y, \Delta_y), y = 1, 2, \ldots, h$. *Sea* $\Delta = E(\Delta_y)$*,* $\boldsymbol{\theta}$ *una matriz semiortogonal y sea* $(\boldsymbol{\theta}\ \boldsymbol{\theta}_0) \in \mathbb{R}^{p\times p}$ *una matriz ortogonal.* $\boldsymbol{\theta}^T\mathbf{X}$ *es una reducción lineal suficiente si y sólo si*

$$\begin{aligned} \mu_y &= \mu + \Delta\boldsymbol{\theta}\nu_y \,, \\ \Delta_y &= \Delta + \Delta\boldsymbol{\theta}T_y\boldsymbol{\theta}^T\Delta \,. \end{aligned}$$

con $\mathrm{E}(\nu_y) = 0$ y $\mathrm{E}(T_y) = 0$.

Estimador MLE para $\boldsymbol{\theta}$

El estimador de máxima verosimilitud $\boldsymbol{\theta}_{LAD}$ *que maximiza la función*

$$\mathcal{L}_{\mathsf{LAD}}(\boldsymbol{\theta}) = \mathsf{const} + \frac{n}{2}\log|\boldsymbol{\theta}^T\tilde{\Sigma}\boldsymbol{\theta}| - \frac{1}{2}\sum_y n_y \log|\boldsymbol{\theta}^T\tilde{\Delta}_y\boldsymbol{\theta}|,$$

provee una reducción lineal suficiente mínima $\boldsymbol{\theta}_{LAD}^T\mathbf{X}$.

HLDA bajo el enfoque de suficiencia

Recordemos que en HLDA $\boldsymbol{\Theta}^T\mathbf{X} \sim \mathcal{N}(\mu_y^*, \Delta_y^*)$

$$\mu_y^* = \begin{pmatrix} \boldsymbol{\theta}^T\mu_y \\ \boldsymbol{\theta}_0^T\mu \end{pmatrix} \qquad \Delta_y^* = \begin{pmatrix} \Omega_y & \mathbf{0} \\ \mathbf{0} & \Omega_0 \end{pmatrix}$$

$$\mathbf{X}|Y = y \sim \mathcal{N}(\mu + \boldsymbol{\theta}\gamma_y, \boldsymbol{\theta}\Omega_y\boldsymbol{\theta}^T + \boldsymbol{\theta}_0\Omega_0\boldsymbol{\theta}_0^T).$$

Si $\Omega = \mathrm{E}(\Omega_y)$, $\Delta = \boldsymbol{\theta}\Omega\boldsymbol{\theta}^T + \boldsymbol{\theta}_0\Omega_0\boldsymbol{\theta}_0^T$. Entonces $\Delta\boldsymbol{\theta} = \boldsymbol{\theta}\mathbf{A}$ y

- $\mu_y = \mu + \boldsymbol{\theta}\gamma_y = \mu + \boldsymbol{\theta}\mathbf{A}\nu_y = \mu + \Delta\boldsymbol{\theta}\nu_y$.
- $\Delta_y - \Delta = \boldsymbol{\theta}(\Omega_y - \Omega)\boldsymbol{\theta}^T = \boldsymbol{\theta}\mathbf{A}T_y\mathbf{A}^T\boldsymbol{\theta}^T = \Delta\boldsymbol{\theta}T_y\boldsymbol{\theta}^T\Delta$.

Luego, $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{HLDA}}$ es una reducción suficiente bajo este modelo para $X|Y = y$.

HLDA bajo el enfoque de suficiencia

Recordemos que en HLDA $\boldsymbol{\Theta}^T\mathbf{X} \sim \mathcal{N}(\mu_y^*, \Delta_y^*)$

$$\mu_y^* = \begin{pmatrix} \boldsymbol{\theta}^T \mu_y \\ \boldsymbol{\theta}_0^T \mu \end{pmatrix} \qquad \Delta_y^* = \begin{pmatrix} \Omega_y & \mathbf{0} \\ \mathbf{0} & \Omega_0 \end{pmatrix}$$

$$\mathbf{X}|Y = y \sim \mathcal{N}(\mu + \boldsymbol{\theta}\gamma_y, \boldsymbol{\theta}\Omega_y\boldsymbol{\theta}^T + \boldsymbol{\theta}_0\Omega_0\boldsymbol{\theta}_0^T).$$

Si $\Omega = \mathrm{E}(\Omega_y)$, $\Delta = \boldsymbol{\theta}\Omega\boldsymbol{\theta}^T + \boldsymbol{\theta}_0\Omega_0\boldsymbol{\theta}_0^T$. Entonces $\Delta\boldsymbol{\theta} = \boldsymbol{\theta}\mathbf{A}$ y

- $\mu_y = \mu + \boldsymbol{\theta}\gamma_y = \mu + \boldsymbol{\theta}\mathbf{A}\nu_y = \mu + \Delta\boldsymbol{\theta}\nu_y$.
- $\Delta_y - \Delta = \boldsymbol{\theta}(\Omega_y - \Omega)\boldsymbol{\theta}^T = \boldsymbol{\theta}\mathbf{A}T_y\mathbf{A}^T\boldsymbol{\theta}^T = \Delta\boldsymbol{\theta}T_y\boldsymbol{\theta}^T\Delta$.

Luego, $\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathrm{HLDA}}$ es una reducción suficiente bajo este modelo para $X|Y = y$.

HLDA bajo el enfoque de suficiencia

Recordemos que en HLDA $\boldsymbol{\Theta}^T\mathbf{X} \sim \mathcal{N}(\mu_y^*, \Delta_y^*)$

$$\mu_y^* = \begin{pmatrix} \boldsymbol{\theta}^T\mu_y \\ \boldsymbol{\theta}_0^T\mu \end{pmatrix} \qquad \Delta_y^* = \begin{pmatrix} \Omega_y & \mathbf{0} \\ \mathbf{0} & \Omega_0 \end{pmatrix}$$

$$\mathbf{X}|Y = y \sim \mathcal{N}(\mu + \boldsymbol{\theta}\gamma_y, \boldsymbol{\theta}\Omega_y\boldsymbol{\theta}^T + \boldsymbol{\theta}_0\Omega_0\boldsymbol{\theta}_0^T).$$

Si $\Omega = \mathsf{E}(\Omega_y)$, $\Delta = \boldsymbol{\theta}\Omega\boldsymbol{\theta}^T + \boldsymbol{\theta}_0\Omega_0\boldsymbol{\theta}_0^T$. Entonces $\Delta\boldsymbol{\theta} = \boldsymbol{\theta}\mathbf{A}$ y

- $\mu_y = \mu + \boldsymbol{\theta}\gamma_y = \mu + \boldsymbol{\theta}\mathbf{A}\nu_y = \mu + \Delta\boldsymbol{\theta}\nu_y$.
- $\Delta_y - \Delta = \boldsymbol{\theta}(\Omega_y - \Omega)\boldsymbol{\theta}^T = \boldsymbol{\theta}\mathbf{A}T_y\mathbf{A}^T\boldsymbol{\theta}^T = \Delta\boldsymbol{\theta}T_y\boldsymbol{\theta}^T\Delta$.

Luego, $\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathsf{HLDA}}$ es una reducción suficiente bajo este modelo para $X|Y = y$.

HLDA bajo el enfoque de suficiencia

Recordemos que en HLDA $\mathbf{\Theta}^T\mathbf{X} \sim \mathcal{N}(\mu_y^*, \Delta_y^*)$

$$\mu_y^* = \begin{pmatrix} \boldsymbol{\theta}^T \mu_y \\ \boldsymbol{\theta}_0^T \mu \end{pmatrix} \qquad \Delta_y^* = \begin{pmatrix} \Omega_y & \mathbf{0} \\ \mathbf{0} & \Omega_0 \end{pmatrix}$$

$$\mathbf{X}|Y = y \sim \mathcal{N}(\mu + \boldsymbol{\theta}\gamma_y, \boldsymbol{\theta}\Omega_y\boldsymbol{\theta}^T + \boldsymbol{\theta}_0\Omega_0\boldsymbol{\theta}_0^T).$$

Si $\Omega = \mathsf{E}(\Omega_y)$, $\Delta = \boldsymbol{\theta}\Omega\boldsymbol{\theta}^T + \boldsymbol{\theta}_0\Omega_0\boldsymbol{\theta}_0^T$. Entonces $\Delta\boldsymbol{\theta} = \boldsymbol{\theta}\mathbf{A}$ y

- $\mu_y = \mu + \boldsymbol{\theta}\gamma_y = \mu + \boldsymbol{\theta}\mathbf{A}\nu_y = \mu + \Delta\boldsymbol{\theta}\nu_y$.
- $\Delta_y - \Delta = \boldsymbol{\theta}(\Omega_y - \Omega)\boldsymbol{\theta}^T = \boldsymbol{\theta}\mathbf{A}T_y\mathbf{A}^T\boldsymbol{\theta}^T = \Delta\boldsymbol{\theta}T_y\boldsymbol{\theta}^T\Delta$.

Luego, $\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathsf{HLDA}}$ es una reducción suficiente bajo este modelo para $X|Y = y$.

HLDA bajo el enfoque de suficiencia

Recordemos que en HLDA $\boldsymbol{\Theta}^T\mathbf{X} \sim \mathcal{N}(\mu_y^*, \Delta_y^*)$

$$\mu_y^* = \begin{pmatrix} \boldsymbol{\theta}^T\mu_y \\ \boldsymbol{\theta}_0^T\mu \end{pmatrix} \qquad \Delta_y^* = \begin{pmatrix} \Omega_y & \mathbf{0} \\ \mathbf{0} & \Omega_0 \end{pmatrix}$$

$$\mathbf{X}|Y = y \sim \mathcal{N}(\mu + \boldsymbol{\theta}\gamma_y, \boldsymbol{\theta}\Omega_y\boldsymbol{\theta}^T + \boldsymbol{\theta}_0\Omega_0\boldsymbol{\theta}_0^T).$$

Si $\Omega = \mathrm{E}(\Omega_y)$, $\Delta = \boldsymbol{\theta}\Omega\boldsymbol{\theta}^T + \boldsymbol{\theta}_0\Omega_0\boldsymbol{\theta}_0^T$. Entonces $\Delta\boldsymbol{\theta} = \boldsymbol{\theta}\mathbf{A}$ y

- $\mu_y = \mu + \boldsymbol{\theta}\gamma_y = \mu + \boldsymbol{\theta}\mathbf{A}\nu_y = \mu + \Delta\boldsymbol{\theta}\nu_y$.
- $\Delta_y - \Delta = \boldsymbol{\theta}(\Omega_y - \Omega)\boldsymbol{\theta}^T = \boldsymbol{\theta}\mathbf{A}T_y\mathbf{A}^T\boldsymbol{\theta}^T = \Delta\boldsymbol{\theta}T_y\boldsymbol{\theta}^T\Delta$.

Luego, $\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathrm{HLDA}}$ es una reducción suficiente bajo este modelo para $X|Y = y$.

HLDA y minimalidad

Cuando Δ_y satisface el modelo de HLDA, el subespacio de reducción estimado, ¿es el menor posible? **No necesariamente**.

Sea $\boldsymbol{\alpha} \in \mathbb{R}^u$ una base para ese subespacio de reducción mínimo.

span$(\boldsymbol{\theta}) \supseteq$ span$(\boldsymbol{\alpha})$ también es un subespacio de reducción suficiente.

En HLDA, $\boldsymbol{\theta}$ es una base para el subespacio que contiene a span(α) tal que $\Delta_y = \boldsymbol{\theta}\Omega_y\boldsymbol{\theta}^T + \boldsymbol{\theta}_0\Omega_0\boldsymbol{\theta}_0^T$.

Existe $\mathbf{A} \in \mathbb{R}^{d\times u}$: $\boldsymbol{\alpha} = \boldsymbol{\theta}\mathbf{A}$ y α maximiza

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}, \mathbf{A}) \quad = \quad & \text{const} - \frac{n}{2}\log|\boldsymbol{\theta}^T\tilde{\Sigma}^{-1}\boldsymbol{\theta}| - \frac{n}{2}\log|\boldsymbol{\theta}^T\tilde{\Sigma}\boldsymbol{\theta}| - \\
& -\frac{n}{2}\log|\mathbf{A}^T\boldsymbol{\theta}^T\tilde{\Sigma}\boldsymbol{\theta}\mathbf{A}| - \frac{1}{2}\sum_y n_y \log|\mathbf{A}^T\boldsymbol{\theta}^T\tilde{\Delta}_y\boldsymbol{\theta}\mathbf{A}|.
\end{aligned}
$$

HLDA y minimalidad

Cuando Δ_y satisface el modelo de HLDA, el subespacio de reducción estimado, ¿es el menor posible? **No necesariamente**.

Sea $\boldsymbol{\alpha} \in \mathbb{R}^u$ una base para ese subespacio de reducción mínimo.

span$(\boldsymbol{\theta}) \supseteq$ span$(\boldsymbol{\alpha})$ también es un subespacio de reducción suficiente.

En HLDA, $\boldsymbol{\theta}$ es una base para el subespacio que contiene a span$(\boldsymbol{\alpha})$ tal que $\Delta_y = \boldsymbol{\theta}\Omega_y\boldsymbol{\theta}^T + \boldsymbol{\theta}_0\Omega_0\boldsymbol{\theta}_0^T$.

Existe $\mathbf{A} \in \mathbb{R}^{d\times u} : \boldsymbol{\alpha} = \boldsymbol{\theta}\mathbf{A}$ y α maximiza

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \mathbf{A}) &= \text{const} - \frac{n}{2}\log|\boldsymbol{\theta}^T\tilde{\Sigma}^{-1}\boldsymbol{\theta}| - \frac{n}{2}\log|\boldsymbol{\theta}^T\tilde{\Sigma}\boldsymbol{\theta}| - \\ &\quad -\frac{n}{2}\log|\mathbf{A}^T\boldsymbol{\theta}^T\tilde{\Sigma}\boldsymbol{\theta}\mathbf{A}| - \frac{1}{2}\sum_y n_y \log|\mathbf{A}^T\boldsymbol{\theta}^T\tilde{\Delta}_y\boldsymbol{\theta}\mathbf{A}|.\end{aligned}$$

HLDA y minimalidad

Cuando Δ_y satisface el modelo de HLDA, el subespacio de reducción estimado, ¿es el menor posible? **No necesariamente**.

Sea $\boldsymbol{\alpha} \in \mathbb{R}^u$ una base para ese subespacio de reducción mínimo.

$\operatorname{span}(\boldsymbol{\theta}) \supseteq \operatorname{span}(\boldsymbol{\alpha})$ también es un subespacio de reducción suficiente.

En HLDA, $\boldsymbol{\theta}$ es una base para el subespacio que contiene a $\operatorname{span}(\alpha)$ tal que $\Delta_y = \boldsymbol{\theta}\Omega_y\boldsymbol{\theta}^T + \boldsymbol{\theta}_0\Omega_0\boldsymbol{\theta}_0^T$.

Existe $\mathbf{A} \in \mathbb{R}^{d\times u}$: $\boldsymbol{\alpha} = \boldsymbol{\theta}\mathbf{A}$ y α maximiza

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}, \mathbf{A}) &= \text{const} - \frac{n}{2}\log|\boldsymbol{\theta}^T\tilde{\Sigma}^{-1}\boldsymbol{\theta}| - \frac{n}{2}\log|\boldsymbol{\theta}^T\tilde{\Sigma}\boldsymbol{\theta}| - \\
& \quad -\frac{n}{2}\log|\mathbf{A}^T\boldsymbol{\theta}^T\tilde{\Sigma}\boldsymbol{\theta}\mathbf{A}| - \frac{1}{2}\sum_y n_y \log|\mathbf{A}^T\boldsymbol{\theta}^T\tilde{\Delta}_y\boldsymbol{\theta}\mathbf{A}|.
\end{aligned}
$$

HLDA y minimalidad

Cuando Δ_y satisface el modelo de HLDA, el subespacio de reducción estimado, ¿es el menor posible? **No necesariamente**.

Sea $\boldsymbol{\alpha} \in \mathbb{R}^u$ una base para ese subespacio de reducción mínimo.

span$(\boldsymbol{\theta}) \supseteq$ span$(\boldsymbol{\alpha})$ también es un subespacio de reducción suficiente.

En HLDA, $\boldsymbol{\theta}$ es una base para el subespacio que contiene a span$(\boldsymbol{\alpha})$ tal que $\Delta_y = \boldsymbol{\theta}\Omega_y\boldsymbol{\theta}^T + \boldsymbol{\theta}_0\Omega_0\boldsymbol{\theta}_0^T$.

Existe $\mathbf{A} \in \mathbb{R}^{d\times u} : \boldsymbol{\alpha} = \boldsymbol{\theta}\mathbf{A}$ y α maximiza

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \mathbf{A}) &= \text{const} - \frac{n}{2}\log|\boldsymbol{\theta}^T\tilde{\Sigma}^{-1}\boldsymbol{\theta}| - \frac{n}{2}\log|\boldsymbol{\theta}^T\tilde{\Sigma}\boldsymbol{\theta}| - \\ &\quad -\frac{n}{2}\log|\mathbf{A}^T\boldsymbol{\theta}^T\tilde{\Sigma}\boldsymbol{\theta}\mathbf{A}| - \frac{1}{2}\sum_y n_y\log|\mathbf{A}^T\boldsymbol{\theta}^T\tilde{\Delta}_y\boldsymbol{\theta}\mathbf{A}|.\end{aligned}$$

HLDA y minimalidad

Cuando Δ_y satisface el modelo de HLDA, el subespacio de reducción estimado, ¿es el menor posible? **No necesariamente**.

Sea $\boldsymbol{\alpha} \in \mathbb{R}^u$ una base para ese subespacio de reducción mínimo.

span$(\boldsymbol{\theta}) \supseteq$ span$(\boldsymbol{\alpha})$ también es un subespacio de reducción suficiente.

En HLDA, $\boldsymbol{\theta}$ es una base para el subespacio que contiene a span$(\boldsymbol{\alpha})$ tal que $\Delta_y = \boldsymbol{\theta}\Omega_y\boldsymbol{\theta}^T + \boldsymbol{\theta}_0\Omega_0\boldsymbol{\theta}_0^T$.

Existe $\mathbf{A} \in \mathbb{R}^{d\times u}$: $\boldsymbol{\alpha} = \boldsymbol{\theta}\mathbf{A}$ y α maximiza

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \mathbf{A}) &= \text{const} - \frac{n}{2}\log|\boldsymbol{\theta}^T\tilde{\Sigma}^{-1}\boldsymbol{\theta}| - \frac{n}{2}\log|\boldsymbol{\theta}^T\tilde{\Sigma}\boldsymbol{\theta}| - \\ &\quad -\frac{n}{2}\log|\mathbf{A}^T\boldsymbol{\theta}^T\tilde{\Sigma}\boldsymbol{\theta}\mathbf{A}| - \frac{1}{2}\sum_y n_y \log|\mathbf{A}^T\boldsymbol{\theta}^T\tilde{\Delta}_y\boldsymbol{\theta}\mathbf{A}|.\end{aligned}$$

Consideraciones

Qué puede heredar HLDA del enfoque de suficiencia?

- Herramientas mejor justificadas para inferir d.
- Eficiencia de cómputo.

Limitaciones

- fijado d, HLDA estima un subespacio de reducción suficiente sólo para una estructura particular de Δ_y.
- $\mathcal{S}_{\text{HLDA}}$ no es invariante ni equivariante ante una transformación de las características $\mathbf{X}$.

Propaganda: ¡¡deberíamos usar θ_{LAD}!!

Consideraciones

Qué puede heredar HLDA del enfoque de suficiencia?

- Herramientas mejor justificadas para inferir d.
- Eficiencia de cómputo.

Limitaciones

- fijado d, HLDA estima un subespacio de reducción suficiente sólo para una estructura particular de Δ_y.
- $\mathcal{S}_{\mathrm{HLDA}}$ no es invariante ni equivariante ante una transformación de las características $\mathbf{X}$.

Propaganda: ¡¡deberíamos usar θ_{LAD}!!

Consideraciones

Qué puede heredar HLDA del enfoque de suficiencia?

- Herramientas mejor justificadas para inferir d.
- Eficiencia de cómputo.

Limitaciones

- fijado d, HLDA estima un subespacio de reducción suficiente sólo para una estructura particular de Δ_y.
- $\mathcal{S}_{\text{HLDA}}$ no es invariante ni equivariante ante una transformación de las características $\mathbf{X}$.

Propaganda: ¡¡deberíamos usar θ_{LAD}!!

Consideraciones

Qué puede heredar HLDA del enfoque de suficiencia?

- Herramientas mejor justificadas para inferir d.
- Eficiencia de cómputo.

Limitaciones

- fijado d, HLDA estima un subespacio de reducción suficiente sólo para una estructura particular de Δ_y.
- $\mathcal{S}_{\text{HLDA}}$ no es invariante ni equivariante ante una transformación de las características $\mathbf{X}$.

Propaganda: ¡¡deberíamos usar θ_{LAD}!!

Consideraciones

Qué puede heredar HLDA del enfoque de suficiencia?

- Herramientas mejor justificadas para inferir d.
- Eficiencia de cómputo.

Limitaciones

- fijado d, HLDA estima un subespacio de reducción suficiente sólo para una estructura particular de Δ_y.
- $\mathcal{S}_{\text{HLDA}}$ no es invariante ni equivariante ante una transformación de las características $\mathbf{X}$.

Propaganda: ¡¡deberíamos usar $\boldsymbol{\theta}_{\text{LAD}}$!!

Consideraciones

Qué puede heredar HLDA del enfoque de suficiencia?

- Herramientas mejor justificadas para inferir d.
- Eficiencia de cómputo.

Limitaciones

- fijado d, HLDA estima un subespacio de reducción suficiente sólo para una estructura particular de Δ_y.
- $\mathcal{S}_{\mathrm{HLDA}}$ no es invariante ni equivariante ante una transformación de las características $\mathbf{X}$.

Propaganda: ¡¡deberíamos usar $\boldsymbol{\theta}_{\mathrm{LAD}}$!!

Consideraciones

Qué puede heredar HLDA del enfoque de suficiencia?

- Herramientas mejor justificadas para inferir d.
- Eficiencia de cómputo.

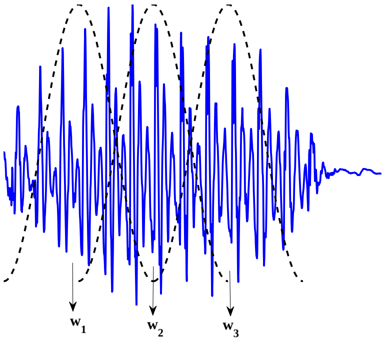
Limitaciones

- fijado d, HLDA estima un subespacio de reducción suficiente sólo para una estructura particular de Δ_y.
- $\mathcal{S}_{\mathsf{HLDA}}$ no es invariante ni equivariante ante una transformación de las características $\mathbf{X}$.

Propaganda: ¡¡deberíamos usar $\boldsymbol{\theta}_{\mathsf{LAD}}$!!

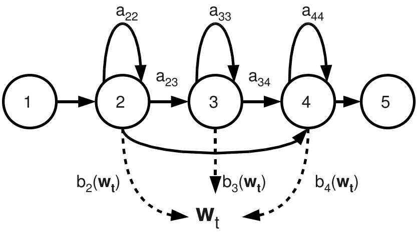
¿Cómo extendemos esto a modelos ocultos de Markov?

Datos secuenciales y modelos ocultos de Markov (HMM)



Datos secuenciales y modelos ocultos de Markov (HMM)

- Las características observadas no son independientes, sino que forman una **secuencia** $\mathbf{X} = \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N$.
- Modelamos esas secuencias mediante **modelos ocultos de Markov**.



Reducción suficiente de dimensiones para HMM

¿Cómo son los modelos de las clases ahora?

$$p(\mathbf{X}|\mathcal{Y} = \vartheta) \quad = \quad \sum_{\mathbf{q}} \prod_{t=1}^{T} b_{q_t}(\mathbf{x_t}) a_{q_{t-1} q_t},$$

Reducción suficiente de dimensiones para HMM

¿Cómo son los modelos de las clases ahora?

$$p(\mathbf{X}|\mathcal{Y}=\vartheta) \quad = \quad \sum_{\mathbf{q}} \prod_{t=1}^{T} b_{q_t}(\mathbf{x_t}) a_{q_{t-1}q_t},$$

¿Cómo estimamos parámetros?

$$\begin{aligned} \mathcal{Q}(\vartheta, \vartheta^{old}) &= \sum_{\mathbf{q}} p(\mathbf{q}|\mathbf{X}, \vartheta^{old}) \log p(\mathbf{q}, \mathbf{X}|\vartheta) \\ &= \sum_{\mathbf{q}} p(\mathbf{q}|\mathbf{X}, \vartheta^{old}) \times \left\{ \sum_t \log a_{q_{t-1}q_t} + \sum_t \log b_{q_t}(\mathbf{x}_t) \right\} \\ &= \mathcal{Q}_a(\vartheta, \vartheta^{old}) + \mathcal{Q}_b(\vartheta, \vartheta^{old}) \end{aligned}$$

Reducción de dimensiones en HMM

Si $b_y(\mathbf{w}_t) = \mathcal{N}(\mathbf{w}_t|\mu_y, \Delta_y)$

$$\begin{aligned} \mathcal{Q}_b(\vartheta, \vartheta^{old}) &= \sum_{y=1}^{N_q}\sum_{t=1}^{T} p(q_t = y|\mathbf{X}, \vartheta^{old}) \log \mathcal{N}(\mathbf{w}_t|\mu_y, \Delta_y) \\ &= \sum_{y=1}^{N_q}\sum_{t=1}^{T} \gamma_t(y) \log \mathcal{N}(\mathbf{w}_t|\mu_y, \Delta_y), \end{aligned}$$

Organización

Simulaciones

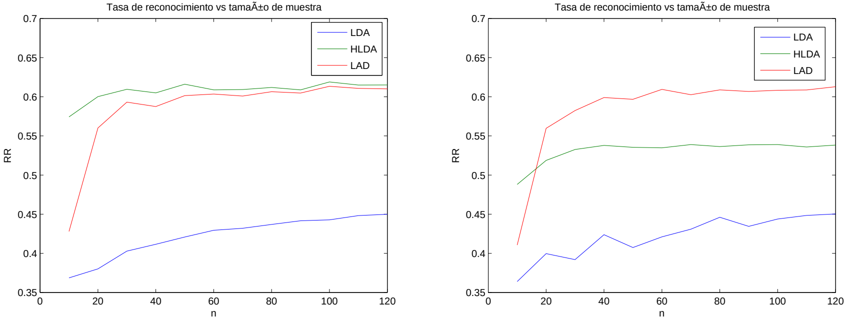


Figura: Tasas de reconocimiento para d conocida. Izq.: datos generados con el modelo de HLDA. Der.: datos generados con el modelo de HLDA y luego multiplicados con una matriz A.

Simulaciones

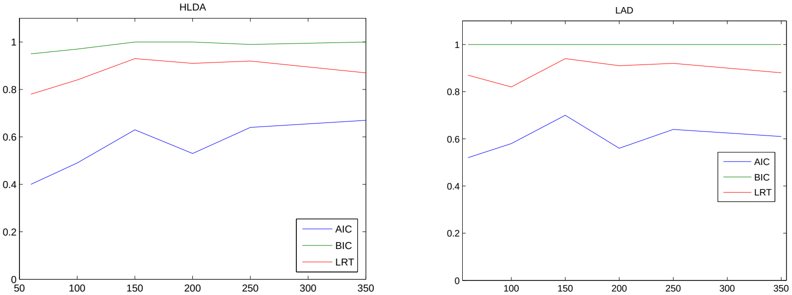


Figura: Inferencia sobre d cuando los datos corresponden al modelo de HLDA.

Simulaciones

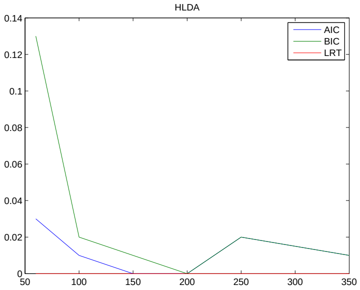


Figura: Inferencia sobre d para los mismo datos, después de multiplicarlos por una matriz $\mathbf{A}$.

Organización

muchas gracias