Redes Bayesianas: Distribución de probabilidad

Aritz Pérez[1] Borja Calvo[2]

Basque Center for Applied Mathematics

UPV/EHU

Donostia, Febrero de 2015

Probabilidad conjunta

$$p(\mathbf{X}_V) = p(X_1, \ldots, X_n)$$

- $\mathbf{X}_V$ ($\mathbf{X}$) es una variable **multinomial**
- Suma **1**: $\sum_{\mathbf{x}_V} p(\mathbf{x}_V) = 1$
- p toma valores en el intervalo $[0, 1]$: $\Omega_V \mapsto [0, 1]$
- Número de parámetros libres **exponencial** en $|V|$:

$$|\Omega_V| - 1 = (\prod_{i \in V} |\Omega_i|) - 1 = (\prod_{i \in V} r_i) - 1$$

Probabilidad marginal

$$p(\mathbf{X}_A) = \sum_{\mathbf{x}_{V \setminus A}} p(\mathbf{X}_A, \mathbf{x}_{V \setminus A})$$

- Se obtiene marginalizando (sumando)
- Suma **1**: $\sum_{\mathbf{x}_A} p(\mathbf{x}_A) = 1$
- Valores en el intervalo $[0, 1]$: $\Omega_A \mapsto [0, 1]$
- Número de parámetros libres **exponencial** en $|A|$:

$$|\Omega_A| - 1 = (\prod_{i \in A} r_i) - 1$$

Probabilidad condicionada

Sean A y B dos subconjuntos disjuntos de V

$$p(\mathbf{X}_A|\mathbf{x}_B) = \frac{p(\mathbf{X}_A, \mathbf{x}_B)}{p(\mathbf{x}_B)} = \frac{p(\mathbf{X}_A, \mathbf{x}_B)}{\sum_{\mathbf{x}_A} p(\mathbf{x}_A, \mathbf{x}_B)}$$

- Para todo $\mathbf{x}_B$, $p(\mathbf{X}_A|\mathbf{x}_B)$ es una distribución de probabilidad
- Número de parámetros libres **exponencial** en $|A| + |B|$:

$$(\prod_{i \in A} r_i - 1) \cdot \prod_{j \in B} r_j$$

- $p(\mathbf{X}_A, \mathbf{X}_B) = p(\mathbf{X}_A|\mathbf{X}_B) \cdot p(\mathbf{X}_B) = p(\mathbf{X}_B|\mathbf{X}_A) \cdot p(\mathbf{X}_A)$

Verosimilitud

Defición

Sean q una distribución de probabilidad y $\mathcal{D}$ un conjunto de datos independiente e identicamente distribuido (i.i.d.) conforme a p, la **verosimilitud** de $\mathcal{D}$ bajo la hipotesis q es:

$$L(\mathcal{D}|q) = \prod_{\mathbf{x}_V \in \mathcal{D}} q(\mathbf{x}_v)$$

- Se emplea como **medida de la calidad** de la distrución q.
- Tiende **a cero exponencialmente** conforme el número de instancias N o la dimensionalidad aumentan n
- Emplear el **logaritmo** de la verosimilitud: $LL(\mathcal{D}|q) = log L(\mathcal{D}|q)$
- LL es negativo, escala linealmente con N y n.

Consistencia

Sean $\mathcal{D}$ un conjunto de datos i.i.d. conforme a p de tamaño N y $\hat{p}$ una distribución conjunta con los parámetros que maximizan la verosimilitud de $\mathcal{D}$. Conforme N tiende a infinito $\hat{p}$ **tiende a** p.

- Probabilidad **empírica**
- Maximizan la probabilidad del conjunto de datos
- La maximización de la verosimilitud se emplea para **aprender los parámetros** de distribuciones de probabilidad

Estimación máximo verosímil

Probabilidad marginal

Sea $\mathcal{D} = \{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$ un conjunto de datos (**entrenamiento**) i.i.d.:

$$\hat{p}(\mathbf{x}_A) = \frac{N(\mathbf{x}_A)}{N}$$

$N(\mathbf{x}_A)$ es el número de instancias que toman el valor $\mathbf{x}_A$ en $\mathcal{D}$

Estimación máximo verosímil

Probabilidad condicionada

$$\hat{p}(\mathbf{x}_A|\mathbf{x}_B) = \frac{N(\mathbf{x}_A, \mathbf{x}_B)}{N(\mathbf{x}_B)}$$

- Maximiza la probabilidad del subconjunto de datos que toman el valor $\mathbf{x}_B$

Ajuste y generalización

Sean $\mathcal{D}$ y $\mathcal{T}$ dos conjuntos i.i.d. donde $\mathcal{D}$ se ha empleado para aprender la distribución q.

- La capacidad de **ajuste** es $\frac{LL(\mathcal{D}|q)}{|\mathcal{D}|\cdot n}$
- La capacidad de **generalización** es $\frac{LL(\mathcal{T}|q)}{|\mathcal{T}\cdot n|}$

Ajuste y generalización

- La verosimilitud de una distribución se puede interpretar como **lo bien que explica** un conjunto de datos
- El **ajuste** tiende a aumentar con el **número de parámetros** del modelo y disminuye con el **número de instancias** $|\mathcal{D}|$
- La **generalización** es una estimación de lo bien que **explica instancias no observadas**

Sobreajuste

- Interesados en distribuciones que **maximicen la generalización**
- El ajuste tiende a tomar **valores mayores** que la generalización
- Cuando ambas medidas difieren hay un **sobreajuste**: **desequilibrio** entre el número de parámetros y de intancias

Robusted de la estimación

Robusted

Cuan **sensible** es la estimación de los parámetros a **cambios** en el conjunto de **datos**

- La robusted **aumenta** con el **número de casos** disponibles (consistencia) y **disminuye** con el **número de parámetros**
- Buscar el **equilibrio**
- **Sobreajuste**: Cuando el número de casos es demasiado poco en comparación con el número de parámetros
- **Aproximaciones Bayesianas** al aprendizaje de parámetros, e.g. corrección de Laplace