Аннотация доклада

Локальная информационная геометрия двоичных цепей Маркова высокого порядка и ее приложения

В.А. Волошко

Пусть $\mathrm{MC}(s)$ – семейство Марковских стационарных вероятностных мер (далее, ВМ) порядка $s \in \mathbb{N}_0 = \{0, 1, \dots\}$ на множестве двоичных последовательностей $\{0,1\}^{\mathbb{Z}}$, $\mathrm{MC} ::= \cup_{s\in\mathbb{N}_0}\mathrm{MC}(s)$ – семейство Марковских стационарных ВМ конечного порядка, $\mathbf{U}$ – равномерная ВМ на $\{0,1\}^{\mathbb{Z}}$. Очевидно:

$$\mathbf{U} \in \mathrm{MC}(0) \subset \mathrm{MC}(1) \subset \mathrm{MC}(2) \subset \cdots \subset \mathrm{MC}.$$

Поэтому касательные пространства $\mathrm{TMC}(s)$ точки $\mathbf{U}$ в многообразиях $\mathrm{MC}(s)$ ($\dim(\mathrm{TMC}(s)) = 2^s$) также образуют вложенную систему:

$$\mathrm{TMC}(0) \subset \mathrm{TMC}(1) \subset \mathrm{TMC}(2) \subset \cdots \subset \mathrm{TMC}, \quad \mathrm{TMC} ::= \cup_{s\in\mathbb{N}_0}\mathrm{TMC}(s).$$

Локальную информационную геометрию двоичных цепей Маркова высокого порядка в окрестности равномерной ВМ $\mathbf{U}$ определяет тензор Фишера-Римана – скалярное произведение на бесконечномерном пространстве TMC:

$$\langle\tau_1, \tau_2\rangle_{\mathrm{TMC}} \in \mathbb{R}, \quad \tau_1, \tau_2 \in \mathrm{TMC}.$$

Пусть теперь задана функция (информативный признак) $f : \{0,1\}^r \to \mathbb{R}^d$ от двоичных r-слов. Будем рассматривать статистики сумм информативных признаков по пересекающимся r-граммам наблюдаемой двоичной последовательности $x_1, \dots, x_T \in \{0,1\}$:

$$\langle f\rangle_x = \sum_{t=1}^{T-r+1} f(x_t, x_{t+1}, \dots, x_{t+r-1}). \tag{1}$$

Статистика (1) для гипотезы "H_0: $\{x_t\}$ – РРСП" имеет асимптотическое нормальное распределение с матрицей ковариаций $T \cdot (\Sigma_f + o(1))$, в случае невырожденности которой может быть построена хи-квадрат статистика:

$$S_f(x) = T^{-1}\left(\langle f\rangle_x - \mathbf{E}_{\mathbf{U}}\{f\}\right)' \Sigma_f^{-1} \left(\langle f\rangle_x - \mathbf{E}_{\mathbf{U}}\{f\}\right), \quad \mathbf{E}_{\mathbf{U}}\{f\} = 2^{-r}\sum_{q\in\{0,1\}^r} f(q). \tag{2}$$

Статистики вида (2) неоднозначно задаются информативными признаками f, но однозначно кодируются конечномерными подпространствами $\mathcal{T} \subset \mathrm{TMC}$, а взаимные свойства таких подпространств отображаются на асимптотические свойства взаимных распределений статистик вида (2). В частности, ортогональность подпространств влечет асимптотическую независимость отвечающих им статистик (2) при истинной гипотезе H_0.