Stanford University

CS 329P : Practical Machine Learning (2021 Fall)

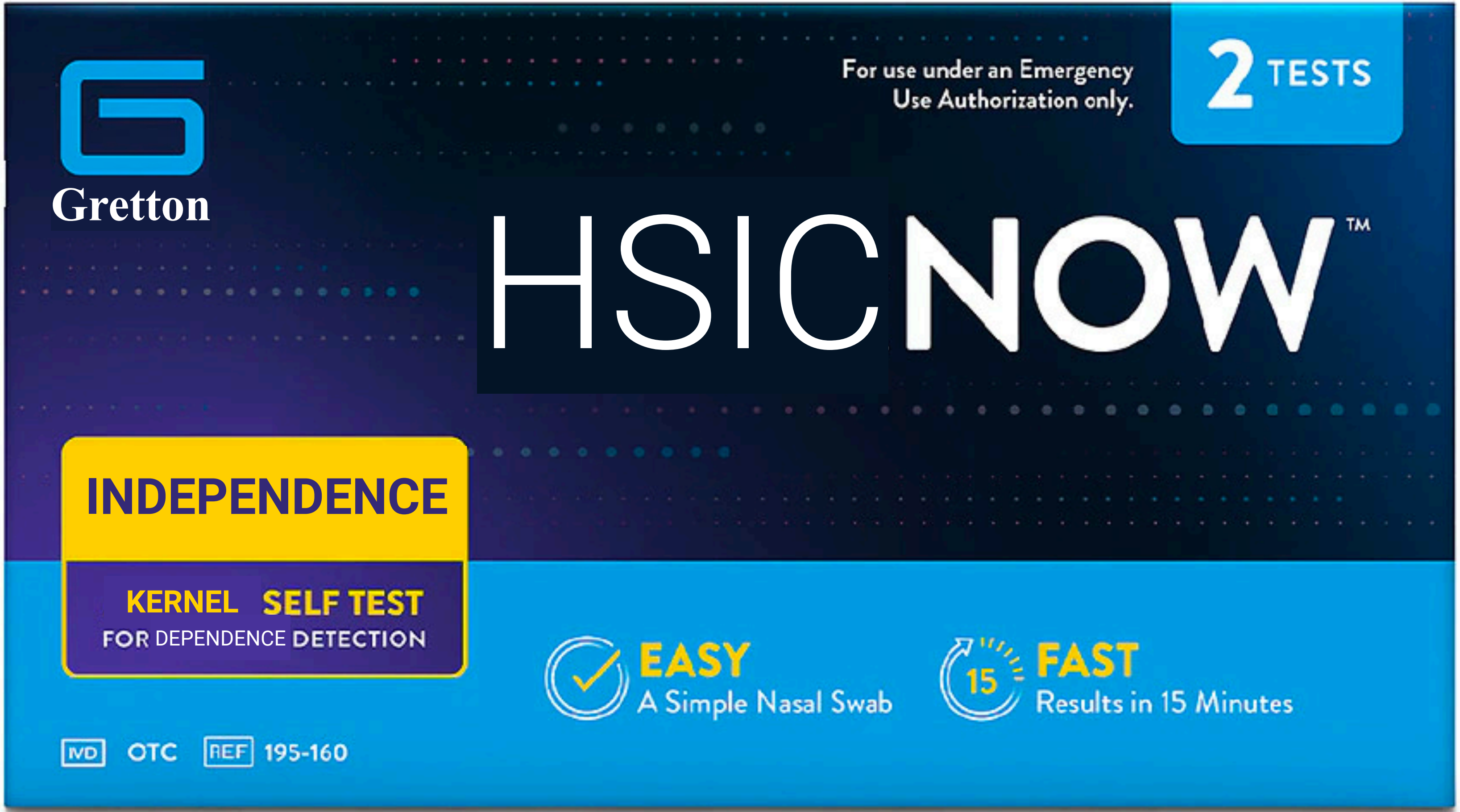
Lecture 8 - Dependent Random Variables

Qingqing Huang, Mu Li, Alex Smola

https://c.d2l.ai/stanford-cs329p

Outline

- **Independence Tests**
- **Sequence Models**
  - Time series & Language
  - Models (Autoregressive, RNNs, Transformers)
- **Graphs**
  - Relational Databases
  - Social Networks
  - Graph Neural Networks


Gretton
For use under an Emergency Use Authorization only.
2 TESTS
HSICNOW™
INDEPENDENCE
KERNEL SELF TEST
FOR DEPENDENCE DETECTION
EASY
A Simple Nasal Swab
15
FAST
Results in 15 Minutes
IVD OTC REF 195-160

Dependent Random Variables $p(x, y) \neq p(x) \cdot p(y)$

Independent Random Variables

$$p(x, y) = p(x) \cdot p(y)$$

Why bother?

- **Dependence**
  - Classification / regression and similar problems need it

$$p(y \mid x) = \frac{p(x, y)}{p(x)}$$

  - A/B testing, cause / effect …
- **Independence** $p(x, y) = p(x) \cdot p(y)$
  - Can ignore variables

$$p(y \mid x) = \frac{p(x, y)}{p(x)} = \frac{p(x) \cdot p(y)}{p(x)} = p(y)$$

Testing for it (classifier)

- Simple discriminative test (classifier) between

$p(x, y)$ and $p(x) \cdot p(y)$

- $Z := \{(x_1, y_1), \ldots (x_n, y_n)\}$ (original data)
- $Z' := \{(x_1, y_{\pi(1)}), \ldots (x_n, y_{\pi(n)})\}$ where π is a random shuffle
- If classifier can distinguish the data, we have dependence.
- Collateral benefit - classifier identifies particularly ‘strongly related’ (read - easy to spot) pairs.

Testing for it (classifier)

- We have a classifier that can tell whether a pair (x, y) is likely to be drawn from a joint distribution via some function $f(x, y)$ since for correct label $f(x, y) > 0$.
- **Crazy idea**
  Use it to build a classifier / regressor by

$$\hat{y} = \underset{y}{\operatorname{argmax}} f(x, y)$$

- **Not so crazy**
  For all 'incorrect' y' we want that $f(x, y') < 0$

Testing for it (MMD)

- **Difference between means**

$$\left\| \mathbf{E}_{(x,y)}[\phi(x)\cdot\phi(y)] - \mathbf{E}_x\mathbf{E}_y[\phi(x)\cdot\phi(y)] \right\|^2$$

Joint expectation

Independent expectation

Can be awkward to compute but needed if the feature maps do not factorize $\phi(x, y)$.

$$f(x,y) = \frac{1}{m}\sum_{i=1}^{m} k(x_i, x) l(y_i, y) - \frac{1}{m^2}\sum_{i=1}^{m} k(x_i, x)\sum_{j=1}^{m} l(y_i, y)$$

Testing for it (HSIC)

- **Covariance operator** (like covariance matrix) should vanish

$$\left\| \mathrm{Cov}_{(x,y)}[\phi(x), \phi(y)] \right\|^2 \left\| \mathbf{E}_{(x,y)} \left[\phi(x) - \mathbf{E}_{x'}[\phi(x')]\right] \cdot \left[\phi(y) - \mathbf{E}_{y'}[\phi(y')]\right] \right\|^2$$

$$= \left\| \mathbf{E}_{(x,y)}[\phi(x) \cdot \phi(y)] - \mathbf{E}_x \mathbf{E}_y[\phi(x) \cdot \phi(y)] \right\|^2$$

Identical term as in MMD. Plenty of algebra yields

$\operatorname{tr} HKHL$ where $H_{ij} = \delta_{ij} - m^{-1}$ and $K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i, y_j)$

Information Theory

- **Kullback Leibler Divergence** (mutual information)
  Compare joint and product of marginals

$$
\begin{aligned}
D(p(x,y)\|p(x)p(y)) &= \int dp(x,y)[\log p(x,y) - \log[p(x)p(y)]] \\
&= H[y] + H[x] - H[(x,y)] \\
&= I(x,y)
\end{aligned}
$$

- Count number of extra bits required to encode X and Y relative to encoding them jointly.
- If the data is independent, no bits can be saved.