3. Averaged One-Dependence Estimators (AODE)

Suppose the data is discrete.

$p(y|\vec{x}) \propto p(y, \vec{x})$

with NB: $\hat{p}(y, \vec{x}) = \left(\prod_{i=1}^{n} \hat{p}(x_i|y)\right) \cdot \hat{p}(y)$.

Can we make weaker independency assumptions?
Yes:

1) LBR:

$$\hat{p}(y, \vec{x}) = \prod_{i=1}^{n} \hat{p}(x_i|y, W) \cdot \hat{p}(y|W),$$

$W = \{x_{j_1}, \dots x_{j_k}\} \subset set(\vec{x})$

a selected set of dependent features.
($W = \emptyset$ gives NB)

2) TAN:

$$\hat{p}(y, \vec{x}) = \prod_{i=1}^{n} \hat{p}(x_i|y, \pi(x_i)) \cdot \hat{p}(y), \text{ where}$$

$\pi(x_i)$ is a "parent" feature for x_i.

~~This gives better~~

These two gives better scores, but worse performance:

	Training Time	Test Time
NB	$O(T \cdot n)$	$O(n \cdot k)$
SP-TAN	$O(T \cdot n^3 \cdot k)$	$O(n \cdot k)$
LBR	$O(T \cdot n)$	$O(n^3 \cdot T \cdot k)$
AODE	$O(T \cdot n^2)$	$O(n^2 \cdot k)$

T - number of training samples,
n - number of features,
k - number of classes.

next step

AODE idea:

$\cancel{p(\vec{x}, f_k) = p}$

$$p(y, \vec{x}) = p(y, x_i) \cdot p(\vec{x} \mid x_i) \quad \forall i = \overline{1,n}$$

$\Downarrow$

$$p(y, \vec{x}) = \frac{\sum_{i \in I_m} p(y, x_i) \cdot p(\vec{x} \mid y, x_i)}{|I_m|}, \text{ where } I_m = \{ i : 1 \le i \le n \text{ and } F(x_i) \ge m \}$$

$F(x_i)$ – # of training examples, containing x_i.

Now we suppose independence and set the following estimation:

$$\hat{p}(y, \vec{x}) = \frac{1}{|I_m|} \cdot \sum_{i \in I_m} \hat{p}(y, x_i) \cdot \prod_{j=1}^{n} \hat{p}(x_j \mid y, x_i),$$

where:

$$\begin{cases} \hat{p}(y) = \dfrac{F(y)+1}{T+d}, \\ \hat{p}(y, x_i) = \dfrac{F(y, x_i)+1}{T_i + d V_i}, \\ \hat{p}(y, x_i, x_j) = \dfrac{F(y, x_i, x_j)+1}{T_{ij} + d V_i V_j} \end{cases}$$

T_i = # of training samples with known class and i-th attribute,

V_i = # of diff. values for attribute i.