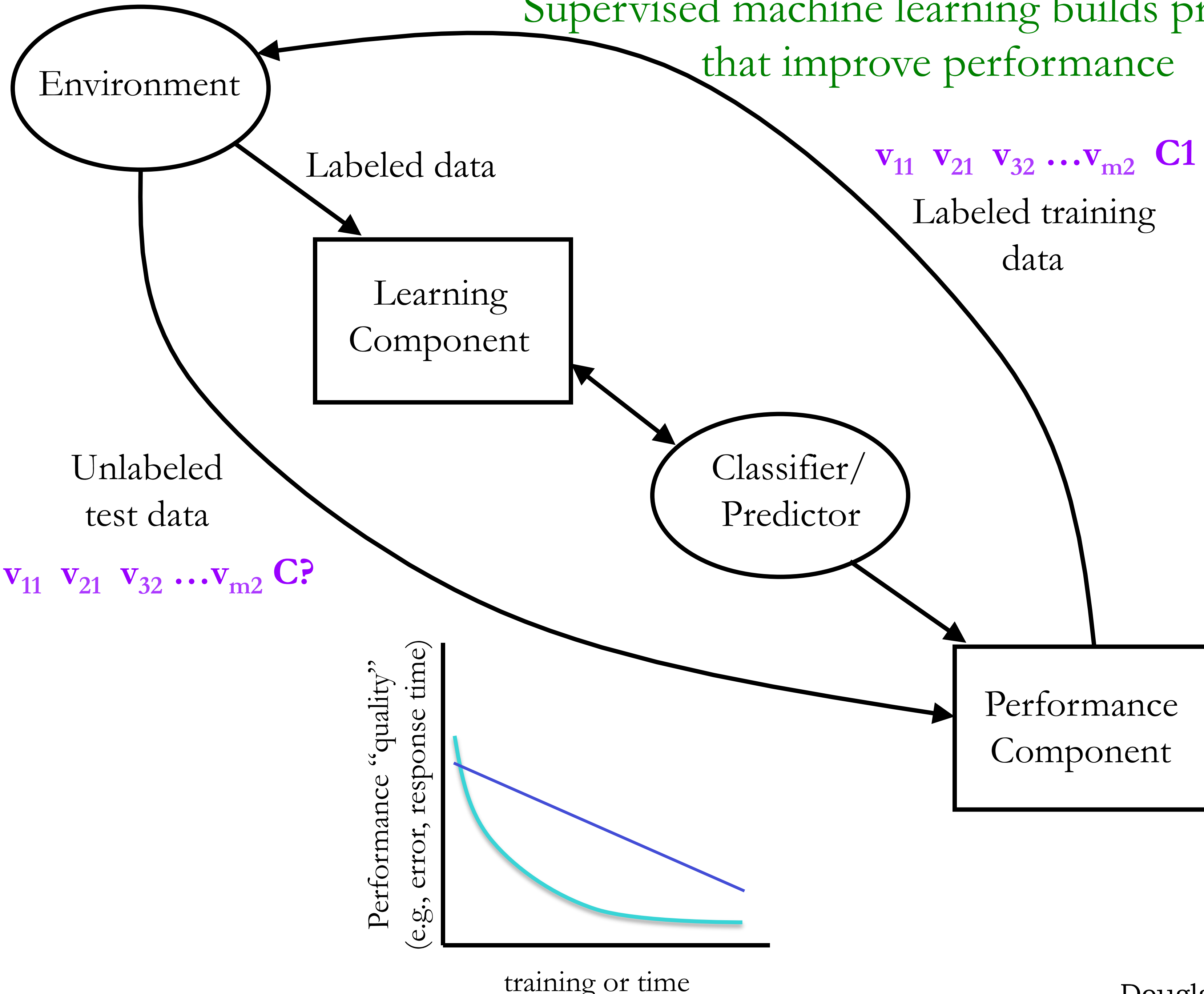
CS 4260 and CS 5260
Vanderbilt University

Lecture on NBCs

This lecture assumes that you have
- Read Section 7.1 through 7.2 of ArtInt and

ArtInt: Poole and Mackworth, Artificial Intelligence 2E
at http://artint.info/2e/html/ArtInt2e.html
to include slides at http://artint.info/2e/slides/ch04/lect1.pdf

Supervised machine learning builds predictors that improve performance



Douglas H. Fisher

Empirical (aka data-driven) Supervised Learning

Given: a set of classified objects (a *training* data set)
Find: a classifier or predictor (for predicting class membership of unclassified data – a *test* set)

An example training set:

Index	V_1	V_2	V_3	. . . .	V_m	C
1	v_{11}	v_{21}	v_{32}		v_{m2}	c_1
2	v_{12}	v_{22}	v_{32}		v_{m1}	c_2
. . . .						
n	v_{12}	v_{21}	v_{31}		v_{m2}	c_1

Subsequent example (but NOT general) assumptions:
two values per variable, two classes

Douglas H. Fisher

Bayesian Classifier

Given a vector $V = \{v_{11}, \quad v_{22}, \quad v_{31}, \quad \ldots, \quad v_{m2}, \quad c_?\}$

Compute:

$$P(c_1 \mid v_{11}, v_{22}, v_{31}, \ldots, v_{m2}) = P(c_1, v_{11}, v_{22}, v_{31}, \ldots, v_{m2}) \, / \, P(v_{11}, v_{22}, v_{31}, \ldots, v_{m2})$$

Compute:

$$P(c_2 \mid v_{11}, v_{22}, v_{31}, \ldots, v_{m2}) = P(c_2, v_{11}, v_{22}, v_{31}, \ldots, v_{m2}) \, / \, P(v_{11}, v_{22}, v_{31}, \ldots, v_{m2})$$

Classify V as in $\mathbf{c_1}$ or $\mathbf{c_2}$, whichever yields higher *conditional* probability

Background: $P(a \mid y) = P(a, y)/P(y)$, where (a, y) is same as $(a \wedge y)$ and $(a \text{ and } y)$

Douglas H. Fisher

Bayesian Classifier cont

Note that denominators are equal, so if we are only interested in most probable, we need only compute the numerators

Compute:

$$P(\mathbf{c_1} \mid v_{11}, v_{22}, v_{31}, \ldots, v_{m2}) \textit{ propto } P(c_1, v_{11}, v_{22}, v_{31}, \ldots, v_{m2}) \cancel{/ P(v_{11}, v_{22}, v_{31}, \ldots, v_{m2})}$$

Compute:

$$P(\mathbf{c_2} \mid v_{11}, v_{22}, v_{31}, \ldots, v_{m2}) \textit{ propto } P(c_2, v_{11}, v_{22}, v_{31}, \ldots, v_{m2}) \cancel{/ P(v_{11}, v_{22}, v_{31}, \ldots, v_{m2})}$$

Classify V as in $\mathbf{c_1}$ or $\mathbf{c_2}$, whichever yields higher *joint* probability

Douglas H. Fisher

Background: the chain rule:

A factorization ordering

$$P(c, v1, v2, v3, v4)$$

$$= \underbrace{\underbrace{\underbrace{\underbrace{P(c) * P(v1 \mid c)}_{P(c,\ v1)} * P(v2 \mid c, v1)}_{P(c,\ v1,v2)} * P(v3 \mid c, v1, v2)}_{P(c,\ v1,\ v2,\ v3)} * P(v4 \mid c, v1, v2, v3)}_{P(c,\ v1,\ v2,\ v3,\ v4)}$$

An alternative ordering

$$P(c, v1, v2, v3, v4)$$

$$= P(v4) * P(v2 \mid v4) * P(v3 \mid v4, v2) * P(v1 \mid v4, v2, v3) * P(c \mid v4, v2, v3, v1)$$

Background: conditional independence

$$P(c, v_1, v_2, v_3, v_4)$$

$$= P(c) * P(v_1 | c) * P(v_2 | c, v_1) * P(v_3 | c, v_1, v_2) * P(v_4 | c, v_1, v_2, v_3)$$

$$= P(c) * P(v_1 | c) * P(v_2 | c) * P(v_3 | c) * P(v_4 | c)$$

if v_is are independent conditioned on c (or conditionally independent given c)

Example: Skin-cover in { hair, feathers, scales }
Heart in {4-chambers, 3-chambers, 2 chambers}
Transport in {walk, fly, swim}
Class in {mammal, bird, fish}

P(mammal, hair, 4-chamber, walk)
≈ P(mammal) * P(hair | mammal) * P(4-chamber | mammal) * P(walk | mammal)
P(fish, scale, 2-chamber, swim) ≈ P(fish) * P(scale | fish) * P(2-chamber | fish) * P(swim | fish)
P(bird, hair, 3-chamber, swim) ≈ P(bird) * P(hair | bird) * P(3-chamber | bird) * P(swim | bird)

Example: Language in {English, Mandarin, Hindi, Spanish, French, …, Tsalagi, Esperanto}
Country in {United States, China, India, Spain, Germany, Mexico, Canada, …}

P(US, English) ≈ P(US)*P(English | US)
P(US, English, Mandarin, French) ≈ P(US) * P(English | US) * P(Mandarin | US) * P(French | US)

Naive Bayesian Classifier

Given a vector $V = \{v_{11}, \quad v_{22}, \quad v_{31}, \quad \ldots, \quad v_{m2}, \quad c_{?}\}$

$P(c_1 | v_{11}, v_{22}, v_{31}, \ldots, v_{m2})$ proportional to $P(c_1, v_{11}, v_{22}, v_{31}, \ldots, v_{m2})$

$$= P(v_{11} | v_{22}, v_{31}, \ldots, v_{m2}, c_1)P(v_{22} | v_{31}, \ldots, v_{m2}, c_1), \ldots, P(v_{m2} | c_1)P(c_1)$$

$$= P(v_{11} | c_1)P(v_{22} | c_1)P(v_{31} | c_1)\ldots.P(v_{m2} | c_1)P(c_1)$$

under assumption that Vi's are independent conditioned on C

$P(c_2 | v_{11}, v_{22}, v_{31}, \ldots, v_{m2})$ proportional to $P(c_2, v_{11}, v_{22}, v_{31}, \ldots, v_{m2})$

$$= P(v_{11} | v_{22}, v_{31}, \ldots, v_{m2}, c_2)P(v_{22} | v_{31}, \ldots, v_{m2}, c_2)\ldots.P(v_{m2} | c_2)P(c_2)$$

$$= P(v_{11} | c_2)P(v_{22} | c_2)P(v_{31} | c_2)\ldots.P(v_{m2} | c_2)P(c_2)$$

under assumption that Vi's are independent conditioned on C

Classify V as in c1 or c2, whichever yields higher joint probability under assumption of conditional independence

Learning a Naive Bayesian Classifier

View probabilities as proportions computed over training set.

$P(v_{11} | c_1) * P(v_{22} | c_1) * P(v_{31} | c_1) * .... * P(v_{m2} | c_1) * P(c_1)$

$= \quad [v_{11}, c_1]/[c1] * [v_{22}, c_1]/[c_1] * [v_{31}, c_1]/[c_1] *...* [v_{m2}, c_1]/[c_1] * [c_1]/[\,]$

where [conditions] is the number of objects/rows in the training set that satisfy all the conditions. So [v11, c1] is the number of training data that are members of c1 and have V1 = v11, [c1] is the number of training objects in c1, [ ] is the total number of training objects.

Learning in this case, is a matter of counting the number of rows in training data in which various conditions satisfied:

- each class/variable-value pair
- each class
- total number of rows

Douglas H. Fisher

$P(v_{11}, v_{22}, v_{31}, c_1)$

$= P(v_{11} \mid v_{22}, v_{31}, c_1) * P(v_{22} \mid v_{31}, c_1) * P(v_{31} \mid c_1) * P(c_1)$

$= [v_{11}, v_{22}, v_{31}, c_1]/[v_{22}, v_{31}, c_1] * [v_{22}, v_{31}, c_1]/[v_{31}, c_1] * [v_{31}, c_1]/[c_1] * [c_1]/[\,]$

$= [v_{11}, v_{22}, v_{31}, c_1] / [\,]$

Given a vector $V = \{1, \quad -1, \quad 0, \quad \ldots., \quad 1\}$

Compute:

There are often unknown values during training (and test)

$P(-1|\ 1, -1, 0, \ldots, 1)$ as $P(1|-1)P(-1|-1)\cancel{P(0|-1)}\ldots.P(1|-1)P(-1)$

$P(1|\ 1, -1, 0, \ldots, 1)$ as $P(1|1)P(-1|1)\cancel{P(0|1)}\ldots.P(1|1)P(1)$

Classify V as in c1 or c2, whichever yields higher probability

$$
\begin{array}{cc}
 & \begin{array}{ccccc} V1 & V2 & V3 & \ldots\ldots\ldots\ldots & Vm \end{array} \\
c1 & \left(\begin{array}{ccccc} [v11,c1] & [v21,c1] & [v31,c1] & \ldots\ldots\ldots\ldots & [vm1,c1] \\ [v12,c1] & [v22,c1] & [v32,c1] & \ldots\ldots\ldots\ldots & [vm2,c1] \end{array}\right) \quad [c1] \\
 & \\
c2 & \left(\begin{array}{ccccc} [v11,c2] & [v21,c2] & [v31,c2] & \ldots\ldots\ldots\ldots & [vm1,c2] \\ [v12,c2] & [v22,c2] & [v32,c2] & \ldots\ldots\ldots\ldots & [vm2,c2] \end{array}\right) \quad [c2] \\
 & \\
 & \begin{array}{ccccc} [v11] & [v21] & [v31] & & [vm1] \\ [v12] & [v22] & [v32] & & [vm2] \end{array} \quad [\,]
\end{array}
$$

Consider an (multidimensional) array implementation of int, and estimate $P(vij|ck)$ as $([vij,ck]+1) / ([ck]+2)$, and $P(ck)$ as $([ck]+1) / ([]+2)$

Number of classes

Number of Vi values

Pseudo-counts to avoid the zero-probability problem and to mitigate overfitting