Lecture 3

Naive Bayes Classification

General concepts, terminology

a machine learning method forms a *hypothesis* about function underlying data

inductive learning: acquiring general concept from specific training examples

main assumption: any hypothesis that approximates target function well over a sufficiently large set of training examples will also approximate target function well over other unobserved examples

generalize: work on unseen examples as well as the training set

need to impose some *inductive bias* – restrict expressivity of hypothesis to

1. make search for hypothesis manageable
2. allow for generalization

characterize different learning methods based on their inductive bias

Curse of Dimensionality

Back to diagnosis example – battery of tests run on patient

recall Bayes Rule:

$$P(c_k|\{A\}) = \frac{P(\{A\}|c_k)P(c_k)}{P(\{A\})}$$

denominator is evidence: probability of observations

if we know priors $P(c_k)$ and likelihoods $P(\{A\}|c_k)$, the most probable class can be determined (can avoid computing evidence)

full MAP approach – classify input $\mathbf{x} = (v_1, v_2, ..., v_N)$:

$$c_{MAP} = \arg\max_{c_k \in C} P(v_1, v_2, ..., v_N|c_k)P(c_k)$$

but this requires knowing for each disease the probability that it will cause any possible combination of symptoms

number of symptom sets exponential in number of basic symptoms (attributes)

Bayesian Diagnosis with Cond. Independence

however if we assume each data attribute (symptom) independent given the diagnosis (class), then:

$$P(\{A\}|c_k) = \prod_{i=1}^{N} P(A_i = v_j|c_k)$$

much simpler to estimate conditional probability of each individual symptom for each possible diagnosis,

can estimate priors, class-conditional distributions based on fraction of training examples in class c_k where $A_i = v_j$ is observed

note: symptoms generally not independent (frequently correlate), so this model not guaranteed to produce reasonable results, but works surprisingly well!

Naive Bayes Classifier

applies to learning tasks:

1. each instance x described by conjunction of attribute values
2. target function takes on value from finite set of classes C

Bayes optimal classifier:

$$c_{MAP} = \operatorname*{arg\,max}_{c_k \in C} P(v_1, v_2, ..., v_N | c_k) P(c_k)$$

Key simplifying assumption: *conditional independence*

$$P(v_1, v_2, ..., v_N | c_k) = \prod_j P(v_j | c_k)$$

Naive Bayes classifier:

$$c_{NB} = \operatorname*{arg\,max}_{c_k \in C} P(c_k) \prod_j P(v_j | c_k)$$

Naive Bayes Classifier: Example

diagnoses: Allergy, Cold, Healthy

symptoms: Sneeze, Cough, Fever

Probability	Well	Cold	Allergy
P(Sneeze$\|D$)	.1	.9	.9
P(Cough$\|D$)	.1	.8	.7
P(Fever$\|D$)	.01	.7	.4

Symptoms: Sneeze=true, Cough=true, Fever=false

compute c_{NB}

Handling Insufficient Data

both prior and conditional probabilities must be estimated from training data, therefore subject to error

if we have few training instances, then the direct probability computation can give probabilities of 0 or 1

example – estimate $P(\mathsf{Cough} = \mathsf{true} | \mathsf{Allergy})$: if value true always observed for feature Cough, then probability will be 1 $\Rightarrow$ no not-coughing person can have an allergy ($P(\mathsf{Allergy} | \mathsf{Cough} = \mathsf{false}) = 0$)

need to smooth estimates to eliminate zeros

Laplace Smoothing

assume binary attribute A, direct estimate:

$$P(A|c_k) = \frac{n_{1k}}{n_{0k} + n_{1k}}$$

Laplace estimate:

$$P(A|c_k) = \frac{n_{1k} + 1}{n_{0k} + n_{1k} + 2}$$

equivalent to prior observation of one example of class k where $A = 0$ and one where $A = 1$

generalized Laplace estimate:

$$P(A_i = v_j|c_k) = \frac{n_{ijk} + 1}{n_k + s_i}$$

- n_{ijk}: number of examples in c_k where $A_i = v_j$
- n_k: number of examples in c_k
- s_i: number of possible values for A_i

Comments on Naive Bayes

- generally works well despite blanket independence assumption
- experiments show it to be quite competitive with other methods on standard datasets
- even when independence assumptions violated, and probability estimates are inaccurate, may still find maximum probability category
- hypothesis constructed directly from parameter estimates derived from training data, no search
- hypothesis not guaranteed to fit training data