20CS6037 Machine Learning

MLE, MAP, Bayesian Reasoning - Chapter 3&5 (Lecture 6: 9/11/14)

Lecturer: Anca Ralescu *Scribes: Khaldoon Ashouiliy, Kyungmook Park*

Section 1: Bayesian Concept Learning

Section 2: The Beta Binomial Model

Section 3: Most Probable Classification

Section 4: The Gamma Distribution

Section 1:
Bayesian Concept Learning

D: Data (set of example for a concept C)
: a point Hypothesis about C.
Note: That both p(D|h) D and can be viewed as functions from the set of instances to {0,1}
C: y $\rightarrow$ {0,1}

$$c(instance) = \begin{cases} 1 & if\ example\ of\ the\ concept\ C \\ 0 & otherwise \end{cases}$$

h and D are consistent if C(i) = h(i) $\forall$ i $\in$ Y

Bayes Theorem
P(h|D) = $\frac{P(h|D)P(h)}{P(D)}$

How to choose hypotheses?

Correct the hypotheses?

- Correct on the training net.
- But not overfitting.

Example Learning a real value function.

f: real valued function.
Training set   D = $\{(x_i, d_i)|d_i$=f(x_i) + $e_i\}$
i=1,. . . ,m

$e_i \sim$ N (0, σ_i)

$\Rightarrow h_{ML} = \underset{h}{argmin} \sum_{i=1}^{m} [d_i - h(x_i)]^2$

<u>Proof</u>

$$
\begin{aligned}
h_{ML} \quad &= \underset{h \in H}{argmax} P(D|h) \\
&= \underset{h \in H}{argmax} P(d_i, \ldots, d_m|h) \ \underline{\text{ind}} \\
&= \underset{h \in H}{argmax} P(d_1|h) x \ldots x P(d_m|h) \\
&= \underset{h \in H}{argmax} \prod_{i=1}^{m} P(d_i|h) \\
&= \underset{h \in H}{argmax} \log[\prod_{i=1}^{m} P(d_i|h)] \\
&= \underset{h \in H}{argmax} \sum_{i=1}^{m} \log P(d_i|h)
\end{aligned}
$$

If iid N(0,σ^2)=0
Then, d_i iid N(f(x_c),σ^2)
iid = independent and identically distributed

from this point on we need to know the actual distribution of (di/h).

Use $e_i \sim$ N (0, σ_i)

P($d_i|h$) $= \frac{1}{\sigma\sqrt{2\bar{u}}} e^{\frac{(d_i - h(x_i))^2}{2\sigma}}$

$\log P(d_i|h) = -\log(\sigma\sqrt{2\bar{u}}) - \frac{1}{2\sigma}(d_i - h(x_i))^2$

$$
\begin{aligned}
\Rightarrow \ h_{ML} \quad &= \underset{h \in H}{argmax} \sum_{i=1}^{m} [-\log(\sigma\sqrt{2\bar{u}}) - \frac{1}{2\sigma}(d_i - h(x_i))^2] \\
&= \underset{h \in H}{argmax} [-\frac{1}{2\sigma} \sum_{i=1}^{m} (d_i - h(x_i))^2] \\
&= \underset{h \in H}{argmax} \sum_{i=1}^{m} (d_i - h(x_i))^2
\end{aligned}
$$

$(d_i - h(x_i))^2$ is the square error between f(x_i) and h(x_i)

$\boxed{h_{ML} \equiv h_{square\ error}}$ $e_i \sim$ N (0, σ_i)

Section 2:
The Beta Binomial Model

Learning to predict probabilities we want to learn f:X→{0,1}
Define $P_0(x)$ = P(f(x) = 0 ; $P_1(x)$ = P(f(x) = 1) = $(-P_0(x))$

example
X = {x|x is a patient with symptom}

$$f(x) = \begin{cases} 1 & if x serving \\ 0 & otherwise \end{cases}$$

We really want to learn the 'Concept' $P_1(x)$ = P(f(x)=1)
based on the learning data

D = $\{< x_i, d_i >, d_i = 0 \; or \; 1 \; i = 1, \ldots, m\}$

What is P(D|h)?

Assume
x_i, d_i are random variables.
x_i and h are independent.

Claim The general
P$(x_i, d_i|h)$=P$(d_i|h, u_i)$P$(x_i|h)$

Proof
Right Hand Side $= P(d_i|h_i u_i)P(x_i|h) = \frac{P(d_i,h,u_i}{\cancel{P(h,u_i)}}\frac{\cancel{P(u_i,h)}}{P(h)}$
$= P(d_i, x_i|h) =$ Left Hand Side

Because x_i and h are independent $\Rightarrow$
P(D|h) = $\prod_{i=1}^{m} P(x_i, d_i|h) = \prod_{i=1}^{m} P(d_i|h, u_i)P(x_i)$
Now $\underline{\text{P}}(d_i = 1|h, u_i) = h(x_i)$

$$\Rightarrow P(d_i|h, u_i) = \begin{cases} h(u_i) & if\ d_i = 1 \\ 1 - h(x_i) & d_i = 0 \end{cases}$$

$\Rightarrow P(d_i|h(x_i)) = [h(x_i)]^{d_i}[1 - h(x_i)]^{1-d_i}$

$\Rightarrow$ P(D|h) = $\prod_{i=1}^{m}[h(x_i)]^{d_i}[1 - h(x_i)]^{1-d_i}P(x_i)$

$$\begin{aligned} h_{ML} \quad &= \underset{h}{argmax} \prod_{i=1}^{m}[h(x_i)]^{d_i}[1 - h(x_i)]^{1-d_i}P(x_i) \\ &= \underset{h}{argmax} \sum_{i=1}^{m}[d_i h(x_i) + \underbrace{(1 - d_i)[1 - h(x_i)]}_{NegativeCrossEntropy} + \cancel{\log P(x_i)}] \end{aligned}$$

Section 3: Most Probable Classification

Suppose $P(h_1|D) = 0.4$ $P(h_2|D) = 0.3$ $P(h_3|D) = 0.3$
$h_1(x) = +$, $h_2(x) = -$, $h_3(x) = -$
$h_M AP = h_1$; The Most Probable Classification

<u>Bayes Optimal Classifier</u>
$= \underset{v \in V}{argmax} \sum_{h \in H} P(v|h)P(h|D)$

$$v \in \{+, -\}$$

h \ P	P(D|h)	P(-|h)	P(+|h)
h_1	0.4	0	1
h_2	0.3	1	0
h_3	0.3	1	0

$\sum_{i=1}^{3} P(+|h_i)P(h_i|D) = 1 \times 0.4 + 0 \times 0.3 + 0 \times 0.3 = 0.4$
$\sum_{i=1}^{3} P(-|h_i)P(h_i|D) = 0 \times 0.4 + 1 \times 0.3 + 1 \times 0.3 = 0.6$
$\Rightarrow \underset{v \in \{-,+\}}{argmax} \sum_{h \in H} P(v|h)P(h|D) = -$

Section 4: The Gamma Distribution

$X \in \Re+$ Random Variable $\sim G(d > 0, \beta > 0)$
d = Shape, β = rate

$f_{Gamma}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ pdf

Where $\Gamma(t) = \int^{\infty} u^{t-1} e^{-u} du \begin{pmatrix} \Gamma(t+1) = t\Gamma(t) \\ \forall t > 0 \end{pmatrix}$

X $\sim$ Gamma(α,β) $\Rightarrow$ E(X) = $\frac{\alpha}{\beta}$;Var(X)=$\frac{\alpha}{\beta^2}$
Mode(X)=$\frac{\alpha-1}{\beta}$

$f'(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}[(\alpha-1)x^{\alpha-2}e^{-\beta x}x^{\alpha-1}e^{-\beta x}] = 0$
$x^{\alpha-2}e^{-\beta x}[\alpha - 1 - \beta x] \Rightarrow \boxed{x = \frac{\alpha-1}{\beta}}$
E(X)=$\frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^\alpha e^{-\beta x} dx \cdots = \frac{\alpha}{\beta}$

<u>The Beta Distribution</u>
X $\sim$ Beta (x|α, β) = $\frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
E(X) = $\frac{\alpha}{\alpha+\beta}$
M(X) = $\frac{\alpha-1}{\alpha+\beta-2}$
Var(X) = $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$