L18 Learning Probabilistic Models.

So far, we've learned.
- what is a Bayes Net.
- the meaning of a Bayes Net.
- answer probabilistic queries.

Where does a Bayes Net come from?
(structure, parameters)
1. ask an expert.
2. learn it from data.

Hypotheses H: h_1, h_2, h_3, h_4, h_5

Data D:
d_1: 1st candy is lime.
d_2: 2nd candy is lime.
d_3: ⋮

A prior over the hypotheses
$P(H) = [0.1, 0.2, 0.4, 0.2, 0.1]$.

Candy Example
- a bag of candies w/ 2 flavours (cherry & lime)
- same wrapper for both flavours
- sold in bags w/ different ratios.
  - h_1: 100% cherry
  - h_2: 75% cherry
  - h_3: 50% cherry
  - h_4: 25% cherry
  - h_5: 0% cherry

After eating N candies, (c cherries, ℓ limes, $N = c + \ell$)
- What is the flavour ratio of the bag?
- What will be the flavour of the next candy?

1 Bayesian learning.
- calculates the probability of each hypothesis given the data

$$P(h_i|d) = \underbrace{\alpha}_{\text{normalizing constant} = \frac{1}{P(d)}} \underbrace{P(d|h_i)}_{\text{likelihood of data given hypothesis}} \underbrace{P(h_i)}_{\text{hypothesis prior}}$$

~~prior~~ prob of hypothesis given data

$$P(h_1|d) = \alpha P(d|h_1) P(h_1)$$
$$= \alpha * 0^2 * 0.1 = \alpha * 0 = 0\%$$
$$P(h_2|d) = \alpha * 0.25^2 * 0.2 = \alpha\, 0.0125 \cong 3.8\%$$
$$P(h_3|d) = \alpha\, 0.5^2 * 0.4 = \alpha\, 0.1 \cong 30.8\%$$
$$P(h_4|d) = \alpha\, 0.75^2 * 0.2 = \alpha\, 0.1125 \cong 34.6\%$$
$$P(h_5|d) = \alpha\, 1^2 * 0.1 = \alpha\, 0.1 \approx 30.8\%$$

$$\alpha = 1/(0 + 0.0125 + 0.1 + 0.1125 + 0.1) = \frac{1}{0.325}$$

② Bayesian prediction. x: next candy is lime.

$$P(x|d) = \sum_i P(x|d \wedge h_i)\, P(h_i|d)$$
$$= \sum_i P(x|h_i)\, P(h_i|d)$$

The weighted average of the predictions of the individual hypotheses.

$$P(x|d) = 0*0 + 0.25*0.038 + 0.5*0.308 + 0.75*0.346 + 1*0.308 = 73.1\%.$$

Properties:

GOOD
- The Bayesian prediction eventually agrees w/ the true hypothesis.
- optimal: Given the prior, the Bayesian prediction is correct more often than any other prediction.
- no overfitting: prior penalizes complex hypotheses.

Price to pay:
- large or infinite hypothesis space.
- the summation/integration may ~~not~~ be intractable to calculate.

Maximum a posteriori (MAP).

- make a prediction based on the most probable hypothesis

$$h_{MAP} = \underset{h_i}{\arg\max}\, P(h_i|d), \quad P(x|d) \cong P(x|h_{MAP})$$

$$P(x|d) = P(x|h_4) = 75\% \qquad h_{MAP} = h_4.$$

Properties:

GOOD
- Finding h_{MAP} is often much easier than Bayesian prediction.
  (opt prob) (summation/integral).
- No overfitting.

- MAP prediction is less accurate than Bayes prediction. but they converge as data increases.
- Finding h_{MAP} may still be intractable.

$$h_{MAP} = \underset{h}{\arg\max}\, P(h|d)$$

$$= \underset{h}{\arg\max}\, P(h) P(d|h)$$

$$= \underset{h}{\arg\max}\, P(h) \prod_i P(d_i|h) \leftarrow \text{non-linear opt.}$$

can take log to linearize.

$$h_{MAP} = \underset{h}{\arg\max} \left[\log P(h) + \cancel{\log} \sum_i \log P(d_i|h) \right].$$

Maximum Likelihood.

simplify MAP by assuming uniform prior.
$P(h_i) = P(h_j) \ \forall i, j.$

$$h_{MAP} = \underset{h}{\operatorname{argmax}} \underbrace{P(h)}_{\text{constant}} P(d|h)$$

$$h_{ML} = \underset{h}{\operatorname{argmax}} P(d|h) \quad \Big/ \quad h_{MAP} = \underset{h}{\operatorname{argmax}} P(h|d).$$

make prediction based on h_{ML} only

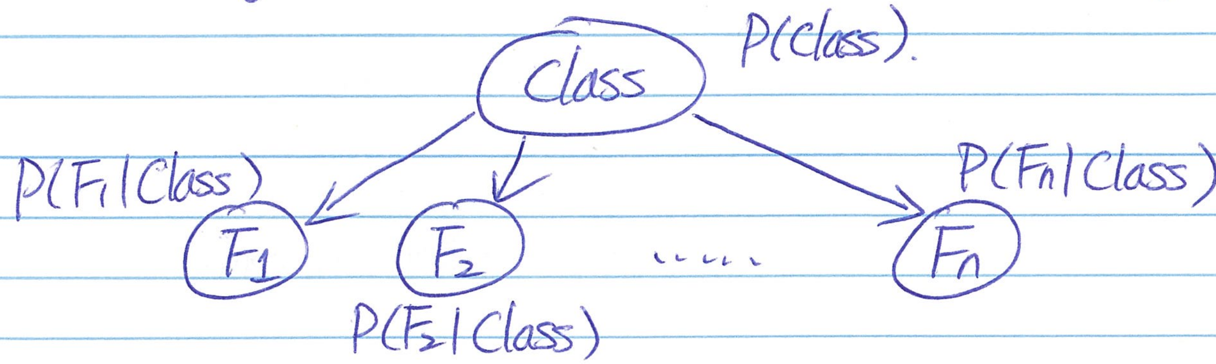
$$P(d|h_1) = 0^2 = 0.$$
$$P(d|h_2) = 0.25^2 = 0.0625.$$
$$\vdots \qquad \vdots$$
$$P(d|h_5) = 1^2 = 1 \qquad h_{ML} = h_5 \quad P(x|h_5) = 1$$

Properties:

GOOD

- h_{ML} is often easier ~~th~~ to find than h_{MAP}.
  $$h_{ML} = \underset{h}{\operatorname{argmax}} \sum_i \log P(d_i|h).$$
- ML prediction is less accurate than Bayesian or MAP. but all converge as data increases.
- susceptible to overfitting.

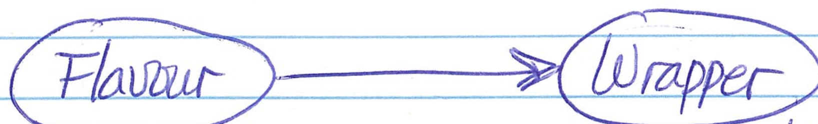
Naive Bayes model – ML Parameter Learning.



"naive": the "feature" variables are not actually conditionally independent given the "class" variables.

– works surprisingly well even when the conditional independence assumption is not true.

Example: red/green wrappers.
– wrapper for each candy is selected probabilistically depending on the flavour.



$P(F=cherry) = \theta$

$P(W=red \mid F=cherry) = \theta_1$

$P(W=red \mid F=lime) = \theta_2$

Unwrap N candies, c cherries, ℓ limes $N = c + \ell$.
cherries: r_c red, g_c green.
lime: r_ℓ red, g_ℓ green.