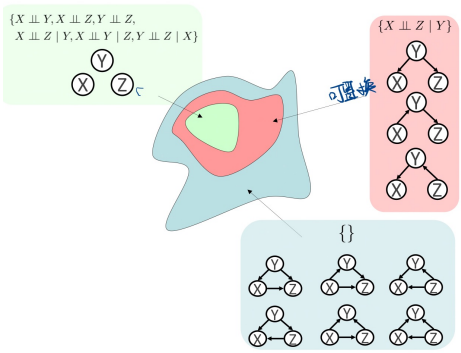
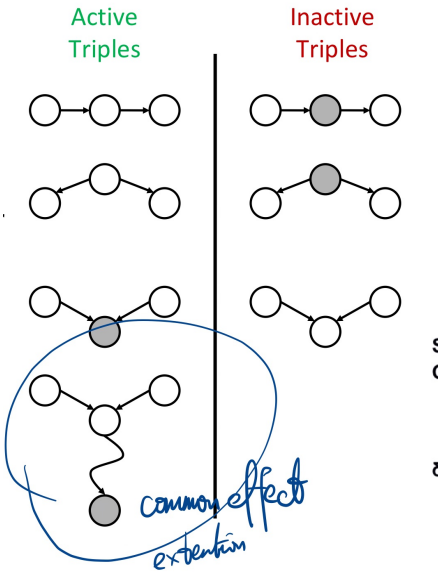
Possibilities

Bayes Rule: $P(x|y) = \frac{P(y|x)}{P(y)} P(x)$





- **Filtering**: $P(X_t|e_{1:t})$
  - **belief state**—input to the decision process of a rational agent
- **Prediction**: $P(X_{t+k}|e_{1:t})$ for $k > 0$
  - evaluation of possible action sequences; like filtering without the evidence
- **Smoothing**: $P(X_k|e_{1:t})$ for $0 \le k < t$
  - better estimate of past states, essential for learning
- **Most likely explanation**: $\arg\max_{x_{1:t}} P(x_{1:t} \mid e_{1:t})$
  - speech recognition, decoding with a noisy channel

State Trellis

value of arc: $P(x_t|x_{t-1})P(e_t|x_t)$

$m_{1:t+1} = \text{VITERBI}(m_{1:t}, e_{t+1})$

$= P(e_{t+1}|X_{t+1}) \max_{x_t} P(X_{t+1}|x_t)\, m_{1:t}$

A: $a_{ji} = P(X_{t+1} = j \mid X_t = i)$

$X_t \mid X_{t+1}$	A	B	H	S
A	0.6	0.1	0.1	0.2
B	0.0	03	0.2	0.5
H	0.8	0.1	0.1	0.1
S	0.2	0.0	0.1	0.7

B: $b_{ik} = P(O_t = k \mid X_t = i)$

$X_t \mid O_t$	p	e	b	l
A	0.6	0.2	0.1	0.1
B	0.1	0.4	0.1	0.4
H	0.0	0.0	0.7	0.3
S	0.2	0.0	0.1	0.9

$\pi = P(X_1 = i)$:

A	B	H	S
0.5	0.0	0.0	0.5

Observations :

$o_{1:4} = \{ b, p, l, e \}$

Step 1: Initialize $\delta_1(i)$

$O_1 =$ b

$\delta_1(i) = \begin{bmatrix} 0.5 \\ 0.0 \\ 0.0 \\ 0.5 \end{bmatrix} \odot \begin{bmatrix} 0.1 \\ 0.1 \\ 0.7 \\ 0.1 \end{bmatrix} = \begin{bmatrix} 0.05 \\ 0 \\ 0 \\ 0.05 \end{bmatrix}$

π $\quad O(l)$

Find:

Most likely hidden state sequence: $X^*_{1:4}$

$O_2 =$ p

$\delta_2(i) =$

		Max probability	Argmax state
max(	0.05 × 0.6 × 0.6 , 0.0 × 0.0 × 0.6 , 0.0 × 0.8 × 0.6 , 0.05 × 0.2 × 0.6)	0.018	A
max(	0.05 × 0.1 × 0.1 , 0.0 × 0.3 × 0.1 , 0.0 × 0.1 × 0.1 , 0.05 × 0.2 × 0.1) =	0.0005	A
max(	0.05 × 0.1 × 0.0 , 0.0 × 0.3 × 0.0 , 0.0 × 0.1 × 0.0 , 0.05 × 0.1 × 0.0)	0	∅
max(	0.05 × 0.7 × 0.0 , 0.0 × 0.5 × 0.7 , 0.0 × 0.1 × 0.0 , 0.05 × 0.7 × 0.0)	0	∅

$O_3 =$ l

$\delta_3(i) =$

		Max probability	Argmax state
max(	0.018 × 0.6 × 0.1 , 0.0005 × 0 × 0.1 , 0 × 0.8 × 0.1 , 0 × 0.2 × 0.1)	0.00108	A
max(	0.018 × 0.1 × 0.4 , 0.0005 × 0.3 × 0.4 , 0 × 0.1 × 0.4 , 0 × 0.0 × 0.4) =	7.2×10^{-4}	A
max(	0.018 × 0.1 × 0.3 , 0.0005 × 0.7 × 0.3 , 0 × 0.1 × 0.1 , 0 × 0.1 × 0.3)	5.4×10^{-4}	A
max(	0.018 × 0.2 × 0.9 , 0.0005 × 0.9 × 0.5 , 0 × 0.1 × 0.9 , 0 × 0.7 × 0.9)	2.8×10^{-4}	A

States and deltas over time:

	$\delta_1(i)$	state	$\delta_2(i)$	state	$\delta_3(i)$	state	$\delta_4(i)$	state
A	0.05	/	0.018	A	0.00108	A	1.296×10^{-4}	A, S
B	0	/	0.0005	A	0.00072	A	8.64×10^{-5}	B
H	0	/	0	/	0.00054	A	0	/
S	0.05	/	0	/	0.00324	A	0	/

Backtracking gives two answers:

$X^*_{1:4} = \{$ A , A , A , A $\}$ and

$X^*_{1:4} = \{$ A , A , S , A $\}$

IN: evidence instantiation

w = 1.0

for i=1, 2, …, n

  if X_i is an evidence variable

    x_i = observation for X_i

    Set w = w * $P(x_i \mid \text{Parents}(X_i))$

  else

    Sample x_i from $P(X_i \mid \text{Parents}(X_i))$

return $(x_1, x_2, \ldots, x_n)$, w

Markov Chain

$P(X_t) = \sum_{x_{t-1}} P(X_{t-1} = x_{t-1}) P(X_t | X_{t-1} = x_{t-1})$

$P(X_{t+1}) = T^T P(X_t)$ ↙ filtering

$P(X_{t+1}|e_{1:t+1}) = \alpha P(e_{t+1}|X_{t+1}) \sum_{x_t} P(x_t|e_{1:t}) P(X_{t+1}|x_t)$

$f_{1:t+1} = \alpha O_{t+1} T^T f_{1:t}$

Prediction: $P(x_{t+1}|e_{1:t}) = \sum_{x_t} P(x_{t+1}|x_t) P(x_t|e_{1:t})$

- X, Y independent if and only if $\forall x, y : P(x, y) = P(x)P(y)$
- X, Y are conditionally independent given Z if and only if: $X \perp\!\!\!\perp Y|Z, \quad \forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$
- Conditional Probability: $P(x|y) = P(x, y)/P(y)$
- Product rule: $P(x, y) = P(x|y)P(y)$
- Chain rule: $P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i|X_1, ..., X_{n-1})$
- Sum rule (marginalization): $p(X) = \sum_y P(X, y)$
- Variant of sum rule: $p(X) = \sum_y P(X|y)P(y)$
- Bayes rule: $P(y|x) = p(x|y)p(y)/P(x)$

- Naive Bayes assumes all features are independent effects of the label.
- Naive Bayes for text: $P(Y, W_1, ..., W_n) = P(Y)\prod_i P(W_i|Y)$, where W_i is the word at position i, $Y \in \{\text{spam, ham}\}$.

Maximum likelihood estimation

- Given the observed set D, find θ to maximize the probability of D
- $\theta = \underset{\theta}{argmax} P(D|\theta)$
- Set the derivative of $P(D|\theta)$ with respect to θ to zero, and solve for θ.

- pretend that we have seen every outcome k extra times.
- $P_{Lap,k} = \frac{c(x)+k}{N+k|X|}$

Linear classifiers

feature values(inputs), weights (learned), activation (sum, $\text{activation}_w(x) = \sum_i w_i \phi_i(x)$)

Binary perceptron learning process

- start with weights=0
- for each training instance (x,y*): classify with current weights, no change if correct else adjust the weight vector by adding or subtracting the feature vector (subtract if y*=-1).

$A \Rightarrow B \equiv \neg A \vee B$

Multiclass perceptron learning process

- start with weights=0
- pick up training examples one by one
- predict with current weights $\hat{y} = argmax_y(w_y \cdot \phi(x))$, no change if correct. Otherwise, lower score of wrong answer and raise score of right answer $w_{\hat{y}} = w_{\hat{y}} - \phi(x)$, $w_{y*} = w_{y*} + \phi(x)$

Probabilistic Perceptron

- softmax function

initialize w (e.g., randomly)
repeat for K iterations:
  for each example (x_i, y_i):
    compute gradient $\Delta_i = -\nabla_w \log P_w(y_i|x_i)$
  compute gradient $\nabla_w \mathcal{L} = \sum_i \Delta_i$
  $w \leftarrow w - \alpha \nabla_w \mathcal{L}$

$$\frac{d}{dw_y} \log P_w(y_i|x_i) = x_i(I(y = y_i) - P(y|x_i))$$

- α: learning rate —- hyperparameter needs to be chosen carefully
- How? Try multiple choices
  - Crude rule of thumb: update should change w by about 0.1-1%
- False ⊨True ✓
- True ⊨ False ✗
- (A ∧ B) ⊨ (A ⇔ B) ✓
- (A ∨ B) ∧ (¬C ∨ ¬D ∨ E) ⊨ (A ∨ B) ✓
- (A ∨ B) ∧ ¬(A ⇒ B) is satisfiable ✓
- (A ⇔ B) ∧ (¬A ∨ B) is satisfiable ✓
- (A ⇔ B) ⇔ C has the same number of models as (A ⇔ B) for any fixed set of proposition symbols that includes A, B, C ✓

3. (5 points) Use Bayes Rule to compute the posterior probabilities $\mathbb{P}(spam \mid \text{“viagra”})$ and $\mathbb{P}(\neg spam \mid \text{“viagra”})$.

Solution: The un-normalized posterior probabilities are:

$\mathbb{P}(spam \mid \text{“viagra”}) = \alpha \cdot \mathbb{P}(\text{“viagra”} \mid spam) \cdot \mathbb{P}(spam) = \alpha \cdot .015 \cdot .60 = \alpha \cdot .0090$
$\mathbb{P}(\neg spam \mid \text{“viagra”}) = \alpha \cdot \mathbb{P}(\text{“viagra”} \mid \neg spam) \cdot \mathbb{P}(\neg spam) = \alpha \cdot .001 \cdot .40 = \alpha \cdot .0004$

Normalizing, so they add up to 1.0, we get: $\mathbb{P}(spam \mid \text{“viagra”}) = 90/94 = 0.9574$
$\mathbb{P}(\neg spam \mid \text{“viagra”}) = 4/94 = 0.0426$.

4.2 Binary Classification

Recall a probabilistic binary classifier $f : \mathcal{X} \to \mathcal{Y}$ is defined as follows:

$$f(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \eta \geq \frac{1}{2} \\ 0 & \text{if } \eta < \frac{1}{2} \end{cases}$$

where $\boldsymbol{x}$ is known to be a non-negative vector and the two classes are encoded 0 and 1. We consider here a probabilistic binary classifier where η is defined by:

$$\eta_{\boldsymbol{\theta}} = \frac{\boldsymbol{w}_1^\top \boldsymbol{x}}{\boldsymbol{w}_0^\top \boldsymbol{x} + \boldsymbol{w}_1^\top \boldsymbol{x}}$$

where $\boldsymbol{\theta} = (\boldsymbol{w}_0, \boldsymbol{w}_1) \in \mathbb{R}_+^{2(D+1)}$ and assuming each instance $\boldsymbol{x} \in \mathbb{R}^{D+1}$ has been appended a constant 1.

1. (3 points) Write the equation of the decision boundary for this classifier.

Solution: $(\boldsymbol{w}_0 - \boldsymbol{w}_1)^\top \boldsymbol{x} = 0$

2. (3 points) What are the possible issues with this model?

Solution: The model defines a linear classifier. The weights may become negative while training, especially if the classification problem is not linearly separable.

3. We would like to learn parameters $\boldsymbol{\theta}$ from a dataset $\mathcal{D} = \{(\boldsymbol{x}^i, y^i) \mid i = 1, \ldots, N\}$, which has been collected in an i.i.d. way.
   (a) (3 points) Write the log-likelihood of $\mathcal{D}$ given parameter $\boldsymbol{\theta}$: $L(\boldsymbol{\theta}; \mathcal{D}) = \log \mathbb{P}(\mathcal{D} \mid \boldsymbol{\theta})$.

Solution: $\mathbb{P}(\mathcal{D} \mid \boldsymbol{\theta}) = \prod_{i=1}^N \left(\frac{\boldsymbol{w}_1^\top \boldsymbol{x}^i}{\boldsymbol{w}_0^\top \boldsymbol{x}^i + \boldsymbol{w}_1^\top \boldsymbol{x}^i}\right)^{y^i} \left(\frac{\boldsymbol{w}_1^\top \boldsymbol{x}^i}{\boldsymbol{w}_0^\top \boldsymbol{x}^i + \boldsymbol{w}_1^\top \boldsymbol{x}^i}\right)^{1-y^i}$
$L(\boldsymbol{\theta}; \mathcal{D}) = \log \mathbb{P}(\mathcal{D}|\boldsymbol{\theta}) = \sum_{i=1}^N y^i \log \boldsymbol{w}_1^\top \boldsymbol{x}^i + (1-y^i) \log \boldsymbol{w}_0^\top \boldsymbol{x}^i - \log\left((\boldsymbol{w}_0 + \boldsymbol{w}_1)^\top \boldsymbol{x}^i\right)$

   (b) (4 points) Write the gradient of this log-likelihood with respect to $\boldsymbol{\theta}$. Recall the gradient is given by:

$$\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathcal{D}) = \left(\left(\frac{\partial L(\boldsymbol{\theta}; \mathcal{D})}{\partial w_{0k}}\right)_{k=1,\ldots,D+1}, \left(\frac{\partial L(\boldsymbol{\theta}; \mathcal{D})}{\partial w_{1k}}\right)_{k=1,\ldots,D+1}\right)$$

Solution: $\frac{\partial L(\boldsymbol{\theta};\mathcal{D})}{\partial w_{1k}} = \sum_{i=1}^N \frac{y^i}{\boldsymbol{w}_1^\top \boldsymbol{x}^i} x_k^i + \frac{1}{(\boldsymbol{w}_0 + \boldsymbol{w}_1)^\top \boldsymbol{x}^i} x_k^i$
$\frac{\partial L(\boldsymbol{\theta};\mathcal{D})}{\partial w_{0k}} = \sum_{i=1}^N \frac{1-y^i}{\boldsymbol{w}_0^\top \boldsymbol{x}^i} x_k^i + \frac{1}{(\boldsymbol{w}_0 + \boldsymbol{w}_1)^\top \boldsymbol{x}^i} x_k^i$

5 General Questions

For each question, answer "yes" or "no" and **justify** your answer.

1. (2 points) For any events E and F, we always have $\mathbb{P}(E \mid F) \geq \min(\mathbb{P}(E), \mathbb{P}(F))$

Solution: (1pt for answer, 1 pt for justification)
No, because if $\mathbb{P}(E) > 0$, $\mathbb{P}(F) > 0$ and $E \cap F = \emptyset$, then $\mathbb{P}(E \cap F) = 0 < \min(\mathbb{P}(E), \mathbb{P}(F))$.

2. (2 points) For any events E and F, we always have $\mathbb{P}(E \mid F) \leq \max(\mathbb{P}(E), \mathbb{P}(F))$

Solution: (1pt for answer, 1 pt for justification)
No, for instance, for $E = F$ and $\mathbb{P}(E) < 1$, we have $\mathbb{P}(E \mid F) = 1$.

3. (2 points) We are given a joint distribution $\mathbb{P}(X, Y)$ over variables X and Y whose domains are respectively $\mathcal{D}_X$ and $\mathcal{D}_Y$. We always have $\sum_{x \in \mathcal{D}_X} \mathbb{P}(X = x \mid Y = y) = 1$ for any $y \in \mathcal{D}_Y$ such that $\mathbb{P}(y) > 0$.

Solution: (1pt for answer, 1 pt for justification)
Yes, $\mathbb{P}(X \mid Y = y)$ is a probability distribution.

4. (2 points) We are given a joint distribution $\mathbb{P}(X, Y)$ over variables X and Y whose domains are respectively $\mathcal{D}_X$ and $\mathcal{D}_Y$. Assume that $\forall y \in \mathcal{D}_Y, \mathbb{P}(y) > 0$. Then, we always have $\sum_{y \in \mathcal{D}_Y} \mathbb{P}(X = x \mid Y = y) = 1$ for any $x \in \mathcal{D}_X$.

Solution: (1pt for answer, 1 pt for justification)
No, $\mathbb{P}(X = x \mid Y)$ is not a probability distribution.