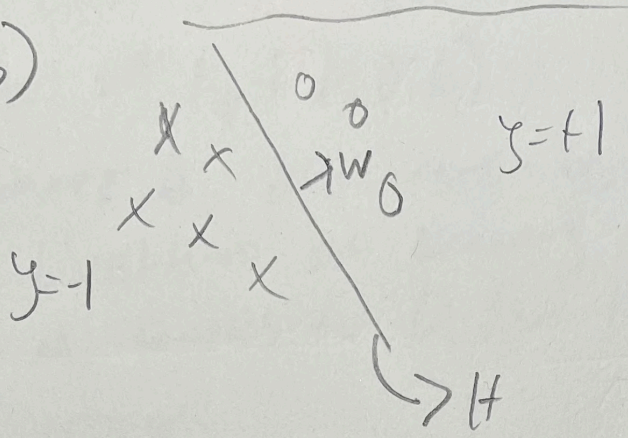
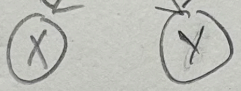
RECAP Lec 02: DECISION BOUNDARY LEARNING

$\hat{y} = \text{sign}(xw^T + b)$

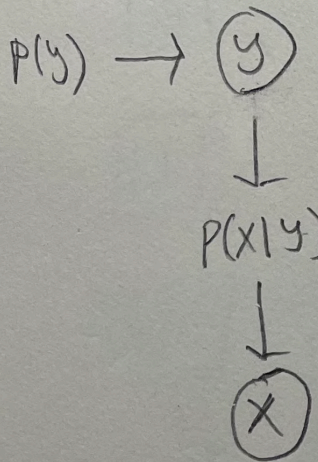


STATISTICAL ML APPROACHES

$P_0(X, Y)$ → REALLY HARD INEFFICIENT!



GENERATIVE MODELING



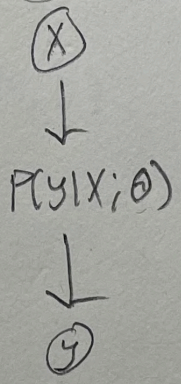
PARAMETER ESTIMATION STEP:
(1) ESTIMATE $P(y; M)$
(2) ESTIMATE $P(x|y; \phi)$

INFERENCE STEP: BAYES RULE

$P(y|x) \propto P(x|y; \hat{\phi}) P(y; \hat{M})$

$\hat{y} = \underset{y}{\text{argmax}}\ P(x|y; \hat{\phi}) P(y; \hat{M})$

DISCRIMINATIVE MODELING



PARAMETER ESTIMATION STEP:
(1) ESTIMATE $P(y|x; \theta)$

INFERENCE:

$\hat{y} = \underset{y}{\text{argmax}}\ P(y|x; \hat{\theta})$

Principle of Maximum Likelihood

- Unobserved data dist: $P_D(D)$ ⇝ $P(y)$, $P(y|x)$, $P(x|y)$
- Approximation of P_D: $P(D;\theta)$
- Goal: estimate, $\hat{\theta}$, that maximizes the likelihood of drawing the M samples in D from $P(D;\theta)$.
- Likelihood function:

$$L(\theta|D) = \prod_{i=1}^{M} P(D^{(i)};\theta)$$

Assumptions: ① All $D^{(i)}$ drawn from P_D
② All $D^{(i)}$ are independent

i.i.d

- Maximum Likelihood Estimation (MLE):

L: $$\hat{\theta} = \underset{\theta}{\text{argmax}} \prod_{i=1}^{M} P(D^{(i)};\theta)$$

LL: $$\hat{\theta} = \underset{\theta}{\text{argmax}} \sum_{i=1}^{M} \log P(D^{(i)};\theta)$$

NLL: $$\hat{\theta} = \underset{\theta}{\text{argmax}} -\sum_{i=1}^{M} \log P(D^{(i)};\theta)$$

↳ $\hat{\theta}$ = ML estimate of θ

GENERATIVE MODELING EXAMPLE: NB

DATA GENERATING PROCESS:

$\forall i \in \{1 \dots m\}$

① $y^{(i)} \sim \text{CATEGORICAL}(\mu)$ — LABEL DIST

② $x^{(i)} \sim \text{MULTINOMIAL}(\phi)$ — FEATURE DIST

EXPLANATION:

- $p(y;\mu) \Rightarrow$ OUR LABEL DIST CAN BE ANY ARBITRARY PMF OVER K CLASSES

$$p(y;\vec{\mu}) = \{\mu_1, \mu_2, \dots, \mu_K\} \qquad \sum_{k=1}^{K} \mu_k = 1$$

INTUITION: SINGLE FLIP OF A K-SIDED COIN

- $p(\vec{x}|y;\phi)$

INTUITION: SUM(X) NUMBER OF FLIPS OF N-SIDED COIN

$$p(\vec{x}|y;\phi) = B(x) \prod_{j=1}^{N} \phi_{kj}^{x_j} \qquad B(x) = \frac{\left(\sum_{j=1}^{N} x_j\right)!}{\prod_{j=1}^{N} x_j!}$$

↳ $p(\vec{x}|y;\phi) = p(x_1 \dots x_n | y;\phi)$

$= p(x_1|y;\phi_1)\, p(x_2|x_1,y;\phi_2) \dots p(x_N|x_1 \dots x_{N-1}, y;\phi_N)$

↑ INTRACTABLE

↳ NB ASSUMPTION:

WORD COUNTS ARE INDEPENDENT CONDITIONED ON SOME CLASS LABEL y.

↳ $p(\vec{x}|y;\phi) = p(x_1|y;\phi)\, p(x_2|y;\phi) \dots p(x_n|y;\phi)$

↳ $\phi \in [0,1]^{K \times N}$, $\forall k \quad \sum_{j=1}^{N} \phi_{kj} = 1$

NB PARAMETER ESTIMATION

- $P(y;\mu) \Rightarrow$ INTUITIVE GUESSTIMATE:

$$\forall_k \quad \hat{\mu}_k = \frac{\text{Number of times } y=k \text{ in } D}{M}$$

$$= \frac{\sum_{i=1}^{M} \delta(y^{(i)}=k)}{M}$$

↳ LETS PROVE THIS W/MLE

$$LL(\mu|D) = \sum_{i=1}^{M} \log \mu_{y^{(i)}} \quad \text{S.T.} \quad \sum_{k=1}^{K} \mu_k = 1$$

↳ $$\hat{\mu} = \underset{\mu}{\text{ARGMAX}} \underbrace{\sum_{i=1}^{M} \log \mu_{y^{(i)}}}_{\ell(\mu)} \quad \text{S.T.} \quad \underbrace{\sum_{k=1}^{K} \mu_k = 1}_{C(\mu)}$$

↳ METHOD of LAGRANGE MULTIPLIERS: $\forall_k \; \nabla_{\mu_k} \ell(\mu) = \lambda \nabla_{\mu_k} C(\mu)$

$$\forall_k \; \nabla_{\mu_k} \left[\sum_{i=1}^{M} \log \mu_{y^{(i)}} \right] = \sum_{i=1}^{M} \delta(y^{(i)}=k) \cdot \frac{1}{\mu_k} = \frac{M_{y=k}}{\mu_k}$$

$$\forall_k \; \nabla_{\mu_k} \left[\sum_{k'=1}^{K} \mu_{k'} \right] = 1$$

↳ $$\forall_k \; \frac{M_{y=k}}{\hat{\mu}_k} = \lambda \Rightarrow \boxed{\forall_k \; \hat{\mu}_k = \frac{M_{y=k}}{\lambda}} \Rightarrow \sum_{k=1}^{K} \hat{\mu}_k = \frac{1}{\lambda} \sum_{k=1}^{K} M_{y=k}$$

↳ $$1 = \frac{M}{\lambda} \Rightarrow \boxed{\lambda = M}$$

↳ $$\boxed{\forall_k \; \hat{\mu}_k = \frac{\text{Count}(y=k)}{M} = \frac{\sum_{i=1}^{M} \delta(y^{(i)}=k)}{M}}$$ ✓

↳ MATCHES GUESSTIMATE!

NB Parameter Estimation Cont.-

↳ Estimate $P(\vec{x}|y;\phi)$ use our Conditional Independence Assumption

↳ $P(\vec{x}|y;\phi) = \prod_{j=1}^{N} P(x_j|y;\phi)$

$$L(\phi|D) = \prod_{i=1}^{M} P(\vec{x}^{(i)}|y^{(i)};\phi) = \prod_{i=1}^{M} B(x) \prod_{j=1}^{N} \phi_{y^{(i)},j}^{x_j^{(i)}}$$

$$LL(\phi|D) = \sum_{i=1}^{M} \log B(x) + \sum_{j=1}^{N} x_j^{(i)} \log \phi_{y^{(i)},j}$$

↳ $\hat{\phi} = \underset{\phi}{\text{Argmax}} \underbrace{LL(\phi|D)}_{\ell(\phi)}$ s.t. $\underbrace{\forall_k \sum_{j=1}^{N} \phi_{kj} = 1}_{C(\phi)}$

↳ Method of Lagrange Multipliers:

$$\forall_k \nabla_{\phi_{kj}} \ell(\phi) = \lambda \nabla_{\phi_{kj}} C(\phi)$$

$$\forall_k \lambda \nabla_{\phi_{kj}} \left(\sum_{j'=1}^{N} \phi_{kj'} \right) = \lambda$$

$$\forall_k \nabla_{\phi_{kj}} \left(\sum_{i=1}^{M} \delta(y^{(i)}=k) \sum_{j'=1}^{N} x_{j'}^{(i)} \log \phi_{kj} \right) = \sum_{i=1}^{M} \delta(y^{(i)}=k) \frac{x_j^{(i)}}{\phi_{kj}}$$

↳ $\forall_{kj} \lambda = \sum_{i=1}^{M} \delta(y^{(i)}=k) \frac{x_j^{(i)}}{\hat{\phi}_{kj}} \Rightarrow \forall_{kj} \lambda \hat{\phi}_{kj} = \sum_{i=1}^{M} \delta(y^{(i)}=k) x_j^{(i)}$

↳ $\forall_k \lambda \sum_{j=1}^{N} \hat{\phi}_{kj} = \sum_{i=1}^{M} \delta(y^{(i)}=k) \sum_{j=1}^{N} x_j^{(i)} \Rightarrow$ But $\forall_k \sum_{j=1}^{N} \phi_{kj} = 1$

↳ $\forall_k \lambda = \sum_{i=1}^{M} \delta(y^{(i)}=k) \sum_{j=1}^{N} x_j^{(i)} \Rightarrow$ λ_k = total number of words in documents with label $y=k$

NB Parameter Estimation Cont...

- $\forall_k \quad \lambda = \sum_{i=1}^{M} \delta(y^{(i)}=k) \sum_{j=1}^{N} x_j^{(i)}$
- From Before we know $\forall_{kj} \quad \sum_{i=1}^{M} \delta(y^{(i)}=k)\, x_j^{(i)} = \lambda \hat{\phi}_{kj}$
- $$\forall_{kj} \quad \hat{\phi}_{kj} = \frac{\sum_{i=1}^{M} \delta(y^{(i)}=k)\, x_j^{(i)}}{\sum_{i=1}^{M} \delta(y^{(i)}=k) \sum_{j'=1}^{N} x_{j'}^{(i)}} = \frac{\text{Count}(\text{class}=k, \text{word}=j)}{\text{Count}(\text{class}=k, \text{all words})}$$

NB INFERENCE:

Now that we have $\hat{\lambda}, \hat{\phi}$, we can compute $P(y|x)$ using BAYES RULE:

$$P(y_{te} \mid x_{te}) \propto P(x_{te} \mid y_{te}; \hat{\phi})\, P(y_{te}; \hat{\lambda})$$

- $$\hat{y}_{te} = \underset{y_{te}}{\text{ARGMAX}} \; \log P(x_{te} \mid y_{te}; \hat{\phi}) + \log P(y_{te}; \hat{\lambda})$$
$$= \underset{k}{\text{ARGMAX}} \sum_{j=1}^{N} x_{te,j} \log \hat{\phi}_{kj} + \log \hat{\lambda}_k$$
$$= \underset{k}{\text{ARGMAX}} \left\{ \vec{x}_{te} (\log \hat{\phi}_{kj})^T + \log \lambda_k \right\}$$

BUT WAIT: Many words will never appear in documents of a given class, therefore many ϕ_{kj} are undefined.

$$\hat{\phi}_{kj} \Leftarrow \frac{\alpha + \text{Count}(k, j)}{\alpha N + \text{Count}(k, \text{all } j)}$$

α can be any positive real number

NB INFERENCE

- Given $\hat{\mu}, \hat{\phi}$: we can now use Bayes Rule to compute the posterior $P(y|x)$:

$$P(y|x) \propto P(x|y;\hat{\phi})\,P(y;\hat{\mu})$$

$$\hat{y}_{Te} = \underset{y_{Te}}{\text{ARGMAX}} \log P(x_{Te}|y_{Te};\hat{\phi}) + \log P(y_{Te};\hat{\mu})$$

$$= \underset{y_{Te}}{\text{ARGMAX}} \sum_{j=1}^{N} x_{Te,j} \log \hat{\phi}_{kj} + \log \hat{\mu}_k$$

$$= \underset{k}{\text{ARGMAX}} \left[\vec{x}_{Te} \cdot (\log \hat{\phi}_k)^T + \log \mu_k \right]$$

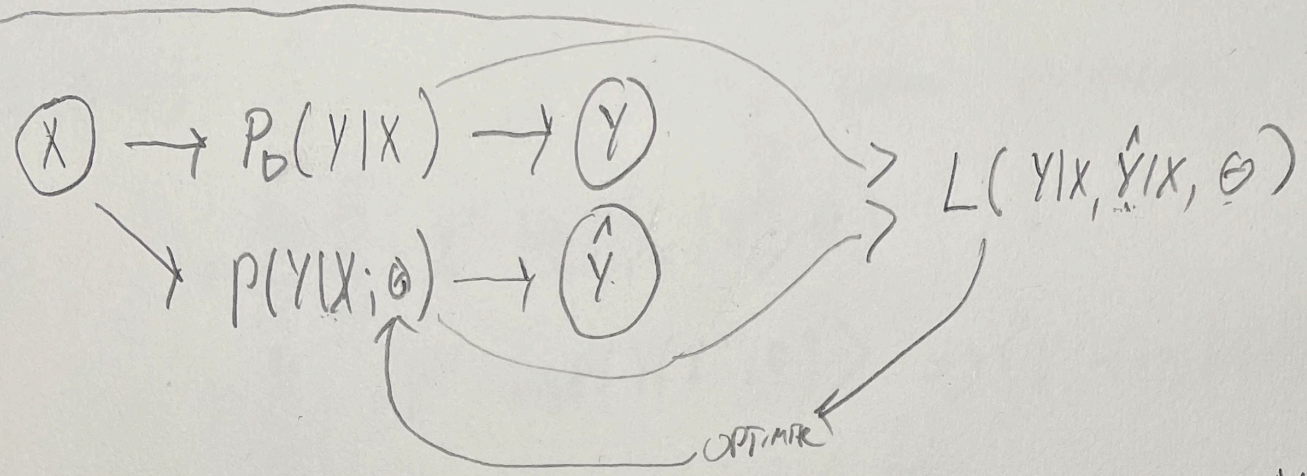
↳ Multinomial NB is a linear classifier

But wait: many k,j combinations will not have support in the data! $\Rightarrow \log 0$

Laplace Smoothing:

$$\hat{\phi}_{kj} = \frac{\alpha + \text{Count}(k,j)}{\alpha \cdot N + \text{Count}(k, \text{all } j)}$$

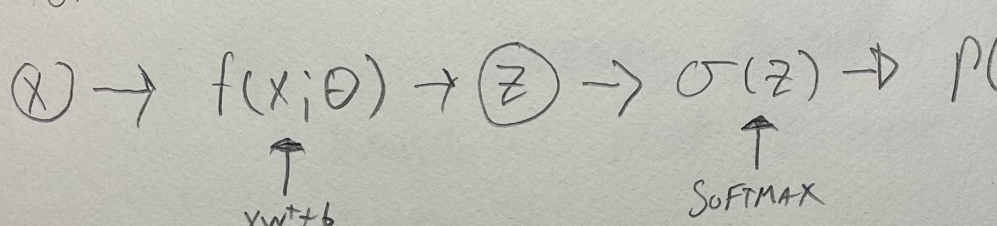
DISCRIMINATIVE LEARNING



$P_D(Y|X)$ = UNOBSERVED DISTRIBUTION THAT GENERATES $Y|X$

$p(Y|X;\theta)$ = APPROX TO P_D PARAMETERIZED BY θ

- GENERAL FRAMEWORK FOR DISCRIMINATIVE CLASSIFIERS:



$f(x;\theta): \mathbb{R}^N \to \mathbb{R}^K$

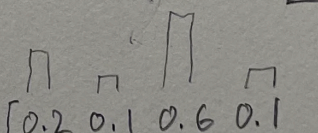
$\sigma(z): \mathbb{R}^K \to [0,1]^K$

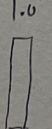
$\sum_{k=1}^{K} \sigma(z_k) = 1$ ← SATISFIES

$\Rightarrow \sigma_{SOFTMAX}(z_k) = \frac{e^{z_k}}{\sum_{k=1}^{K} e^{z_k}}$

$\vec{y} \in \{0,1\}^K \longrightarrow$ SET, NOT A RANGE

$\sum_{k=1}^{K} y_k = 1$





↳ E.g.: $\vec{y} = [0,0,1,0] \in \{0,1\}^4$ $\quad y_4 = 1, y_{k \neq 4} = 0$

MLE IN DISCRIMINATIVE LEARNING

$$NLL(\theta|D) = \sum_{i=1}^{M} -\log P(Y_K \mid X^{(i)}; \theta) \quad \text{where } K = \text{ARGMAX } y^{(i)}$$

$$= \sum_{i=1}^{M} \sum_{K=1}^{K} -y_K^{(i)} \log P(Y_K \mid X^{(i)}; \theta)$$

$$= \sum_{i=1}^{M} -\langle y^{(i)}, \log P(Y \mid X^{(i)}; \theta) \rangle \Rightarrow \langle Y, -\log P(Y \mid X; \theta) \rangle$$

$$= \sum_{i=1}^{M} E_{Y|X \sim P_D}[-\log P(Y \mid X^{(i)}; \theta)]$$

$$= \sum_{i=1}^{M} H(P_D, P(Y \mid X^{(i)}; \theta))$$

↳ CROSS ENTROPY BETWEEN THE OBSERVED EMPIRICAL DIST Y AND OUR MODEL DIST P_θ.

SOFTMAX REGRESSION:

$f(x;\theta) \Rightarrow XW^T + b \Rightarrow$ $\theta = \{W, b\}$

$\sigma(z) \Rightarrow$ Softmax function $W \in \mathbb{R}^{K \times N}$

$b \in \mathbb{R}^K$

$$\sigma(f(x; W, b))_K = \frac{e^{XW_K^T + b_K}}{\sum_{K'=1}^{K} e^{XW_{K'}^T + b_{K'}}}$$

Softmax Regression

$$NLL(\theta|D) = \sum_{i=1}^{m} \sum_{k=1}^{K} -y_k^{(i)} \log P(Y_k | x^{(i)}; \theta)$$

↳ Plug in $\sigma_{SOFTMAX}$:

$$NLL(W,b|D) = -\sum_{i=1}^{m} \sum_{k=1}^{K} y_k^{(i)} \log \frac{e^{x^{(i)} W_k^T + b_k}}{\sum_{k'=1}^{K} e^{x^{(i)} W_{k'}^T + b_{k'}}}$$

↳ $-\sum_{i=1}^{m} \left[\sum_{k=1}^{K} y_k^{(i)} \left(x^{(i)} W_k^T + b_k \right) \right.$ $\left. - \log \left[\sum_{k'=1}^{K} e^{x^{(i)} W_{k'}^T + b_{k'}} \right] \right]$

↳ $-\sum_{i=1}^{m} \left[\left(\sum_{k=1}^{K} y_k^{(i)} \left(x^{(i)} W_k^T + b_k \right) \right) - \log \left[\sum_{k=1}^{K} e^{x^{(i)} W_k^T + b_k} \right] \right]$

↳ **GRADIENT DESCENT:**

1. COMPUT $\nabla_\theta NLL(\theta|D)$
2. UPDATE $\theta = \theta - \eta \nabla_\theta$

↳ $\nabla_{W_k} NLL(W,b|D) \in \mathbb{R}^N$

$$= -\sum_{i=1}^{m} \nabla_{W_k} \left[\sum_{k'=1}^{K} y_{k'}^{(i)} \left(x^{(i)} W_{k'}^T + b_{k'} \right) \right] - \nabla_{W_k} \log \left[\sum_{k'=1}^{K} e^{x^{(i)} W_{k'}^T + b_{k'}} \right]$$

$$= -\sum_{i=1}^{m} y_k^{(i)} x^{(i)} - x^{(i)} \cdot \frac{e^{x^{(i)} W_k^T + b_k}}{\sum_{k'=1}^{K} e^{x^{(i)} W_{k'}^T + b_{k'}}}$$

$$\Rightarrow \boxed{\nabla_{W_k} = \sum_{i=1}^{m} x^{(i)} \left(P(Y_k | x^{(i)}; W, b) - y_k^{(i)} \right)} \Rightarrow \text{GRADIENT DESCENT}$$

$$\boxed{\nabla_{W_k} = x^{(i)} \left(P(Y_k | x^{(i)}; W, b) - y_k^{(i)} \right)} \Rightarrow \text{STOCHASTIC GRADIENT DESCENT}$$

Softmax Regression Continued

$$NLL(W,b|D) = -\sum_{i=1}^{M} y_k^{(i)} - \frac{e^{x^{(i)} W_k^T + b_k}}{\sum_{k'=1}^{K} e^{x^{(i)} W_{k'}^T + b_{k'}}}$$

$$\Rightarrow \nabla_{b_k} NLL(W,b|D) =$$

$$\nabla_{b_k} = -\sum_{i=1}^{M} p(y_k | x_k^{(i)}; W, b) - y_k^{(i)} \Rightarrow \text{GRADIENT DESCENT}$$

$$\nabla_{b_k} = p(y_k | x^{(i)}; W, b) - y_k^{(i)} \Rightarrow \text{STOCHASTIC GRADIENT DESCENT}$$

$\Rightarrow$ MINI BATCH GRADIENT DESCENT:

$1 \leq \text{BATCH-SIZE} < M$