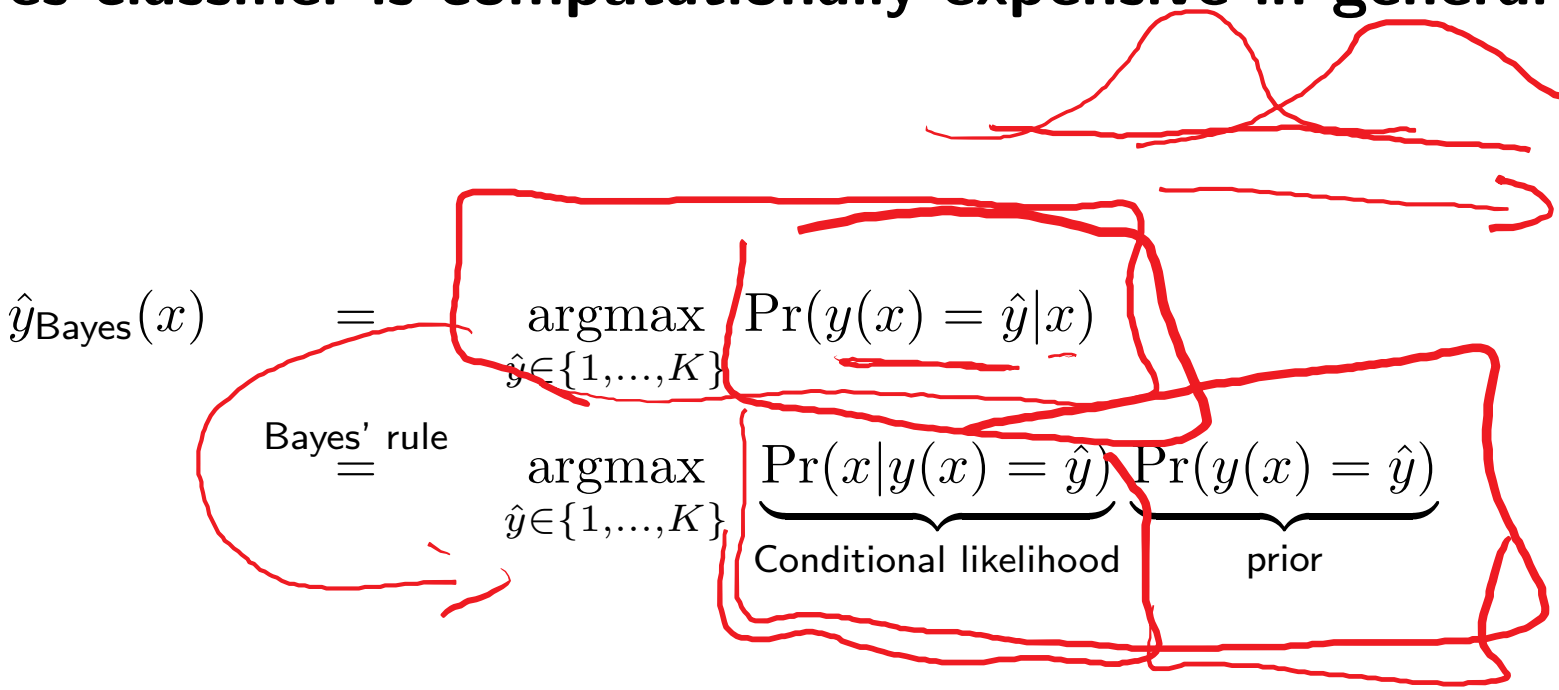
9. Naive Bayes

- Naive Bayes
- Generative vs discriminative models

Many slides borrowed from Prof. Minh Hoai Nguyen's previous course offering

Bayes classifier is computationally expensive in general



$$\hat{y}_{\text{Bayes}}(x) = \operatorname*{argmax}_{\hat{y} \in \{1,\ldots,K\}} \Pr(y(x) = \hat{y} | x)$$

Bayes' rule

$$= \operatorname*{argmax}_{\hat{y} \in \{1,\ldots,K\}} \underbrace{\Pr(x | y(x) = \hat{y})}_{\text{Conditional likelihood}} \underbrace{\Pr(y(x) = \hat{y})}_{\text{prior}}$$

- Assume K classes, D attributes, each taking N values.
- Sampling complexity for $y|x$ is $O(KN^D)$
- Problem: no modeling assumption on $\hat{y}$ = high sampling complexity

Naive Bayes

assumption unnecessary

Assumption: each attribute is independent given label

$$\Pr(x|y) = \prod_k \Pr(x_k|y)$$

Example: Will it rain today?

- is it cold? ⟷ is temperature low?
- is it cloudy?
- are my joints aching?
- did it rain yesterday?

Is this assumption reasonable? (Given it will rain, are cold, cloudy independent?)

Naive Bayes

Assumption: each attribute is independent

$$\Pr(x|y) = \prod_k \Pr(x_k|y)$$

Example: Will it rain today?

- is it cold?
- is it cloudy?
- are my joints aching?
- did it rain yesterday?

Is this assumption reasonable? (Given it will rain, are cold, cloudy independent?)

Maybe not, but it's useful and works fine in practice

"All models are wrong, but some are useful" – George Box

Naive Bayes

$$\hat{y}_{\text{Bayes}}(x) = \operatorname*{argmax}_{\hat{y} \in \{1,\ldots,K\}} \Pr(y(x) = \hat{y} | x)$$

Bayes' rule

$$= \operatorname*{argmax}_{\hat{y} \in \{1,\ldots,K\}} \Pr(x | y(x) = \hat{y}) \Pr(y(x) = \hat{y})$$

Naive Bayes' assumption

$$= \operatorname*{argmax}_{\hat{y} \in \{1,\ldots,K\}} \prod_{i=1}^{D} \Pr(x_i | y(x) = \hat{y}) \Pr(y(x) = \hat{y})$$

From $O(KN^D)$ to $O(KDN)$

Example: Bayes classifier, no structural assumption

$$\hat{y}_{\text{Bayes}}(x) = \underset{\hat{y} \in \{1,\ldots,K\}}{\operatorname{argmax}} \Pr(x|y(x) = \hat{y}) \Pr(y(x) = \hat{y})$$

- Will it rain?
- Collect historical data

rain	yes	no	yes	no	no	yes
A	1	1	-1	1	1	-1
B	99	41	4	18	3	12
C	♢	♡	♠	♣	♣	♡

To see every value of $\Pr(x_A, x_B, x_C|y(x) = \hat{y})$ at least once, require

$$\underbrace{2}_{\#\text{ classes}} \times \underbrace{2}_{\#\text{ values for A}} \times \underbrace{100}_{\#\text{ values for B}} \times \underbrace{4}_{\#\text{ values for C}} \quad \text{observations}$$

Naive Bayes classifier

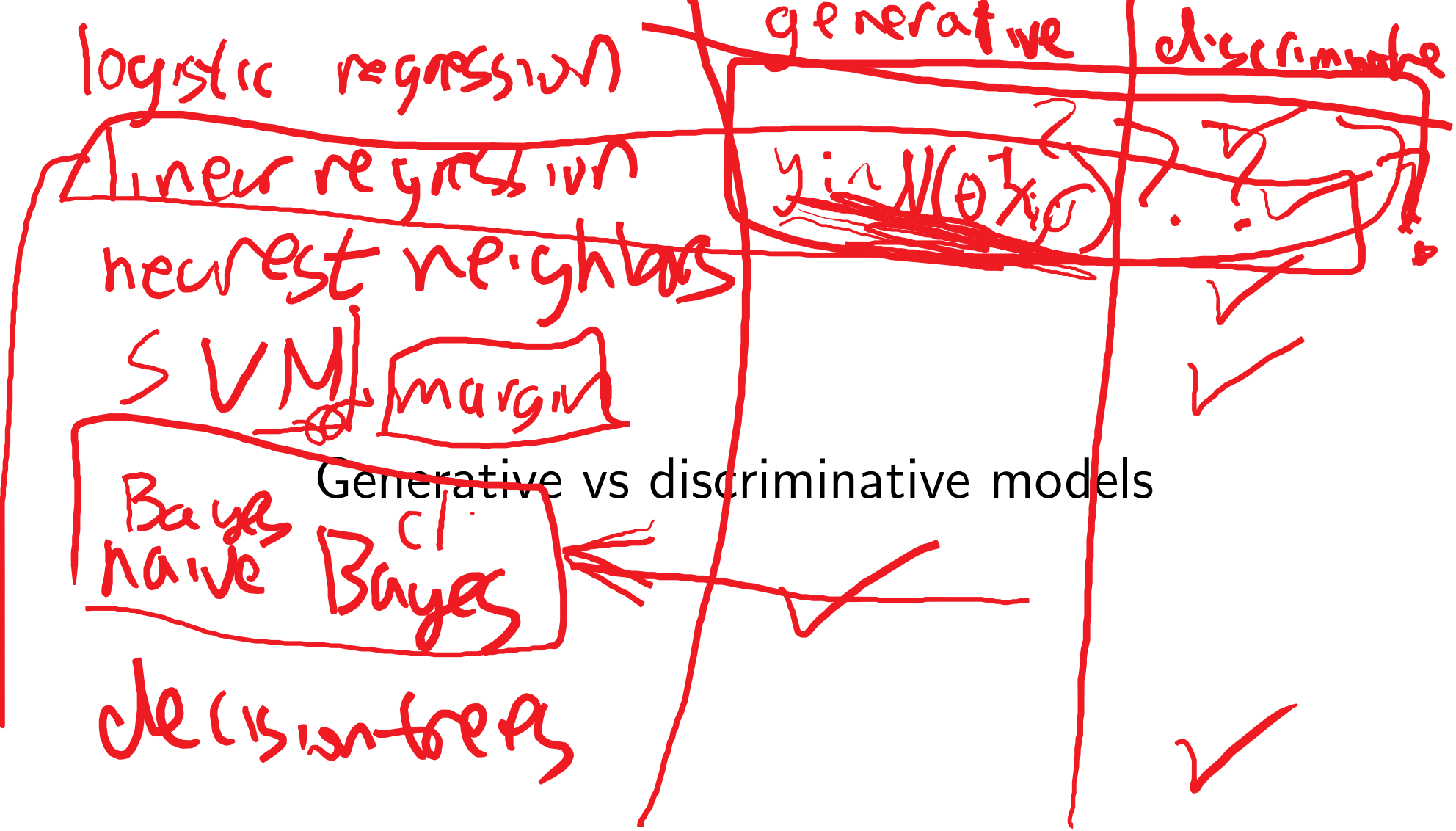
$$\hat{y}_{\text{Bayes}}(x) = \operatorname*{argmax}_{\hat{y} \in \{1,\ldots,K\}} \prod_{i=1}^{D} \Pr(x_i | y(x) = \hat{y}) \Pr(y(x) = \hat{y})$$

- Will it rain?
- Collect historical data

rain	yes	no	yes	no	no	yes
A	1	1	-1	1	1	-1
B	99	41	4	18	3	12
C	♢	♡	♠	♣	♣	♡

To see every value of $\Pr(x_A|y), \Pr(x_B|y), \Pr(x_B|y)$ at least once, require

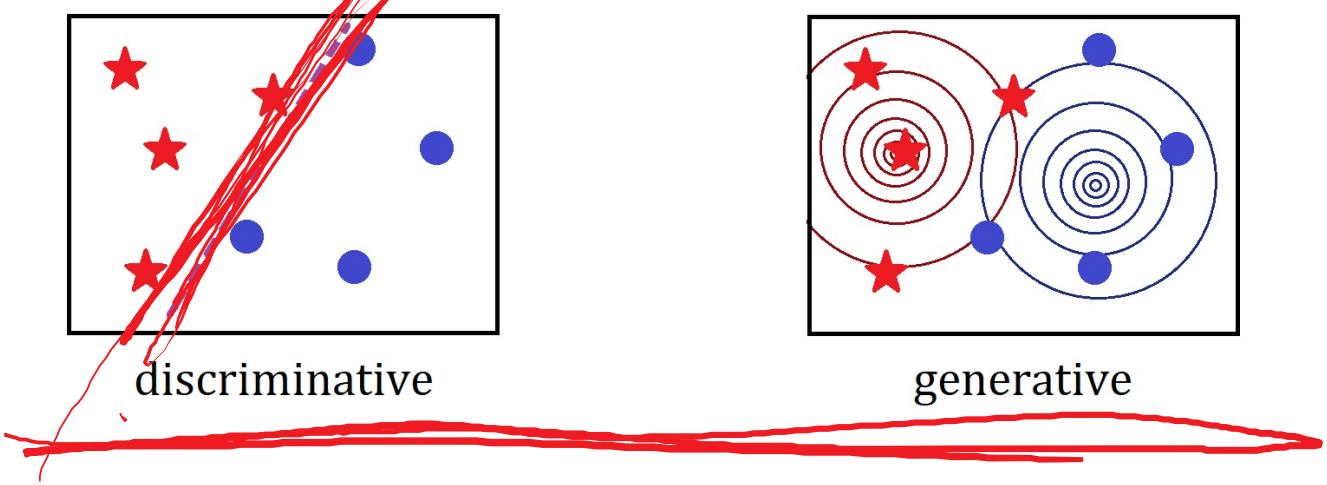
$$\underbrace{2}_{\#\text{ classes}} \times \left(\underbrace{2}_{\#\text{ values for A}} + \underbrace{100}_{\#\text{ values for B}} + \underbrace{4}_{\#\text{ values for C}} \right) \text{ observations}$$



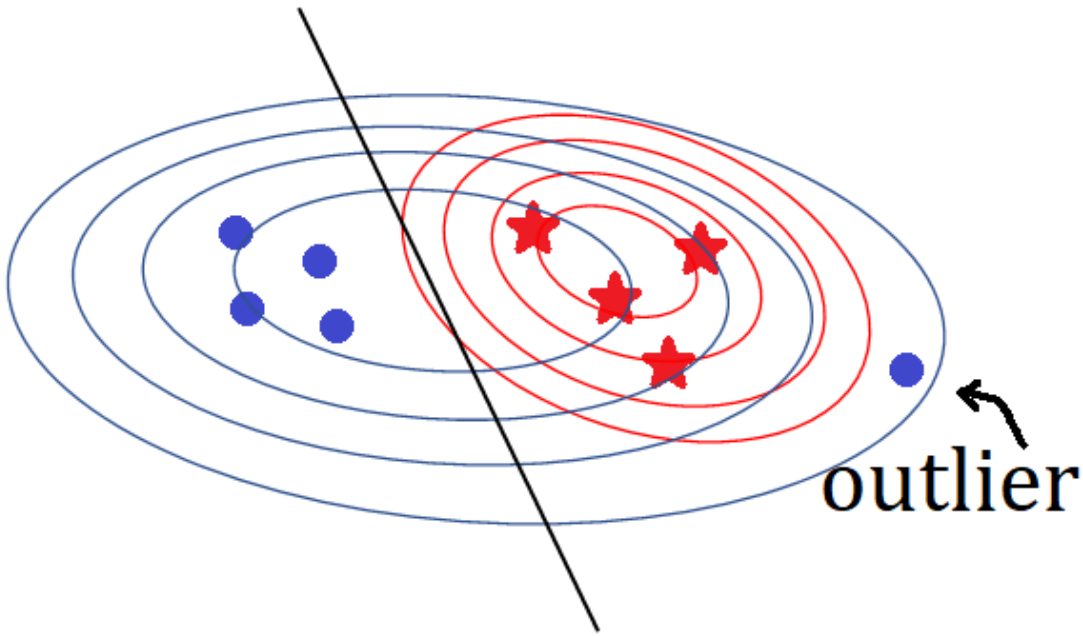
Generative vs discriminative models

Discriminative vs generative models

- Discriminative models focus on task
  - Goal: find the margin that minimizes misclassification rate
  - e.g.: logistic regression, regression trees, K-nearest neighbors
- Generative models focus on interpretability
  - First build a probabilistic model, then find MLE / MAP
  - e.g.: linear regression, naive Bayes, Gaussian mixture models



Generative models may not be robust



Generative vs discriminative models

- In classification, the goal is to produce a rule $\hat{y}(x) \approx y(x) \in \{1, ..., K\}$
- A **generative classifier** aims to model the entire universe
  - Assume a model $\Pr(x|y)$ and prior $\Pr(y)$
  - Produces $\hat{y}(x) = \underset{y \in \{1,...,K\}}{\operatorname{argmax}} \Pr(x|y)\Pr(y)$
  - Requires a lot of training samples
  - Interpretable
- A **discriminative classifier** models only the decision rule
  - Directly models $f(x, k) = \Pr(y = k|x)$ for all k
  - Produces $\hat{y}(x) = \underset{k \in \{1,...,K\}}{\operatorname{argmax}} f(x, k)$
  - Requires much less training samples
  - Hard to interpret

Example: Logistic regression for binary classification

$$\theta^* := \operatorname*{argmax}_{\theta} \sum_{i=1}^{m} \log(y_i x_i^T \theta), \qquad y(x) = \mathbf{sign}(x^T \theta^*)$$

Generative or discriminative?

Example: Logistic regression for binary classification

$$\theta^* := \operatorname*{argmax}_{\theta} \sum_{i=1}^{m} \log(y_i x_i^T \theta), \qquad y(x) = \mathbf{sign}(x^T \theta^*)$$

Generative or discriminative? Ans: Discriminative!

Remember modeling assumption:

$$\log\left(\frac{\Pr(y(x)=1)}{\Pr(y(x)=-1)}\right) = x^T\theta$$

No assumptions on $\Pr(x|y)$!

Example: Linear regression

$$\theta^* := \operatorname*{argmin}_{\theta} \sum_{i=1}^{m} (x_i^T \theta - y_i)^2 + \|\theta\|_2^2, \qquad y(x) = \mathbf{sign}(x^T \theta^*)$$

Generative or discriminative?

Example: Linear regression

$$\theta^* := \operatorname*{argmin}_{\theta} \underbrace{\sum_{i=1}^{m} (x_i^T \theta - y_i)^2 + \|\theta\|_2^2}_{f(\theta)}, \qquad y(x) = \mathbf{sign}(x^T \theta^*)$$

Generative or discriminative? Ans: Generative!

<-- depending on interpretation

Under Gaussian modeling assumption:

$$y \sim \mathcal{N}(\theta^T x, 1), \qquad \theta \sim \mathcal{N}(0, I) \Rightarrow f(\theta) \propto \Pr(\theta | \{x_i, y_i\})$$

Summary

- Up until now, we only considered discriminative classifiers
  - Logistic regression
  - Thresholded linear regression
- Naive Bayes: first generative approach
- Key assumption: each features independent given label
  - Reasonable?
  - Maybe not accurate, but effective
  - Considerably reduces complexity