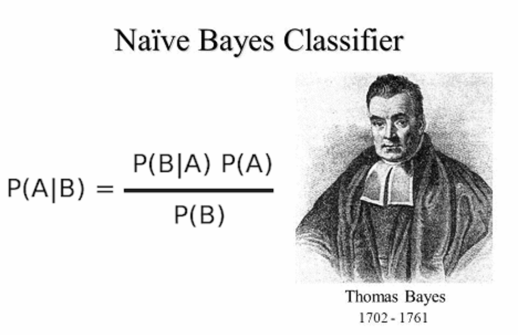
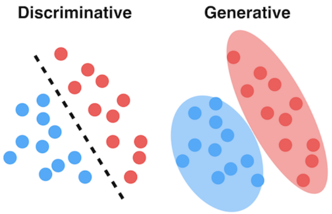
Machine Learning course

Generative modeling for supervised learning

Cédric Vincent-Cuaz
cedric.vincent-cuaz@inria.fr



Discriminative vs Generative Classifiers



Discriminative classifiers

- Look for boundaries between samples $\{\boldsymbol{X}_i\}_{i \in [N]} \subset \mathcal{X}$ to separate classes $\{y_i\}_{i \in [N]} \subset \mathcal{Y}$. *Examples*: k-NN, SVM.
- Directly learn the conditional probability $\mathbb{P}(y|x)$: How does the target variable y behave observing features x ?

Generative classifiers

- Learn the joint distribution $\mathbb{P}(x, y)$: How the features x and the target variable y occur together ? *Examples*: **Naive Bayes classifier**, Hidden Markov Models.
- Under a statistical hypothesis over $\mathbb{P}(x|y)$ perform classification using Bayes' rule.

Probability basics: definitions

Consider two realizations A and B of a random variable

Some probabilities relation

- Prior probability: $\mathbb{P}(A)$
- Conditional probability: $\mathbb{P}(A|B) = \mathbb{P}(A)$ if B is true.
- Joint probability: $\mathbb{P}(A, B) = \mathbb{P}(A \cap B)$
- Relationship: $\mathbb{P}(A, B) = \mathbb{P}(A|B)\mathbb{P}(B)$
- If A and B are independent:
  - $\mathbb{P}(A|B) = \mathbb{P}(A)$ and $\mathbb{P}(B|A) = \mathbb{P}(B)$
  - $\mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B)$.

Probability basics: examples

Two six-sided dices: After rolling both dices denoted D_1 and D_2, we can have the following events:

- (A) $D_1 = 3$
- (B) $D_2 = 1$
- (C) $D_1 + D_2 = 8$

What are the following probabilities?

- $\mathbb{P}(A) = ?$
- $\mathbb{P}(B) = ?$
- $\mathbb{P}(C) = ?$

Two six-sided dices: After rolling both dices denoted D_1 and D_2, we can have the following events:

- (A) $D_1 = 3$ (dice 1 lands on 3)
- (B) $D_2 = 1$ (dice 2 lands on 1)
- (C) $D_1 + D_2 = 8$ (dices sum to 8)

What are the following probabilities?

- $\mathbb{P}(A) = \mathbb{P}(B) = \frac{1}{6}$.
- $\mathbb{P}(C) = \frac{5}{36}$.

$$\begin{aligned}\mathbb{P}(C) &= \mathbb{P}(\cup_{d \in \{2,..,6\}} \{D_1 = d, D_2 = 8 - d\}) \\ &= \sum_{d \in \{2,..,6\}} \mathbb{P}(\{D_1 = d, D_2 = 8 - d\}) \\ &= \sum_{d \in \{2,..,6\}} \mathbb{P}(D_1 = d)\mathbb{P}(D_2 = 8 - d)\end{aligned}$$

Probability basics: examples

Two six-sided dices: After rolling both dices denoted D_1 and D_2, we can have the following events:

- (A) $D_1 = 3$
- (B) $D_2 = 1$
- (C) $D_1 + D_2 = 8$

What are the following probabilities?

- $\mathbb{P}(A, B) = ?$
- $\mathbb{P}(A, C) = ?$

Probability basics: examples

Two six-sided dices: After rolling both dices denoted D_1 and D_2, we can have the following events:

- (A) $D_1 = 3$
- (B) $D_2 = 1$
- (C) $D_1 + D_2 = 8$

What are the following probabilities?

- $\mathbb{P}(A, B) = \mathbb{P}(A|B)\mathbb{P}B = \frac{1}{6} * \frac{1}{6} = \frac{1}{36}$
- $\mathbb{P}(A, C) = \mathbb{P}(C, A) = \mathbb{P}(C|A)\mathbb{P}(A) = \frac{1}{6} * \frac{1}{6} = \frac{1}{36}$
- Nb: $\mathbb{P}(A, C) \neq \mathbb{P}(A)\mathbb{P}(C)$, A and C are not independent.

LIKELIHOOD
The probability of "B" being True, given "A" is True

PRIOR
The probability "A" being True. This is the knowledge.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

POSTERIOR
The probability of "A" being True, given "B" is True

MARGINALIZATION
The probability "B" being True.

Bayes' rule: illness example

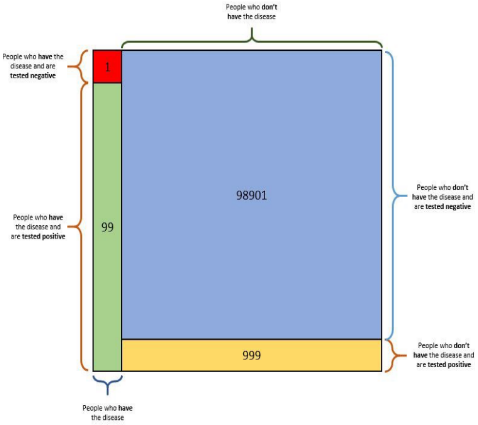
- You wake up one morning with a slight discomfort and decide to visit a doctor.
- The doctor performs few tests to detect a rare disease that only happen to one person out of thousand (1/1000).
- Unfortunately the results are positive.
- The doctor tells you that **the tests identify the disease** with 99% accuracy.
- But, what is the possibility that you may have this disease ?

Bayes' theorem: illness example

- It is a rare disease which concerns only 0.1% of the global population: For 100,000 people,
  - 100 people have the rare disease
  - 99,900 do not !
- The test accuracy is 99%: If 100 people have this rare disease
  - 99 would test positive
  - 1 would test negative !
- What do we forget here ?
  - If people without the disease are also tested:
    - 1% of those tested would also be falsely declared sick: 1 % of 99,900 people $\Longrightarrow$ 999 persons !
  - The probability that you have the disease because you test positive:

$$\frac{\text{number of ill people tested positive}}{\text{total number of people tested positive}} = \frac{99}{99+999} = 9\%$$

Bayes' theorem: illness example



- Hypothesis (H): tested person is ill or not.
- Evidence (E): the test is positive or negative.
- Bayes' theorem can be used to find the probability that you have the disease (H=True) given the evidence provided by a positive test result:

$$\mathbb{P}(H|E) = \frac{\mathbb{P}(E|H)\mathbb{P}(H)}{\mathbb{P}(E)}$$

Bayes' theorem: Rare disease detection

- You are the Chief of staff of the Minister of Health. A disease affects 1 out of 1000 people in the population (0.1 %)
- A manager of a major pharmaceutical company comes to you withh his newscreening test:
    - If a person is sick, the test is 92% positive
    - If a person is not ill, the test is positive at 0.04 %
- These results sound excellent, what do you think ?
- Do you validate the test in order to authorize its commercialization ?

Bayes' theorem: Rare disease detection

Before authorizing the marketing of this test, what should be checked?

- Only the results presented by the laboratory ?
- Or what is the probability that a person is really sick (S) if their test (T) is positive (**true positive**)? $\mathbb{P}(S = 1|T = 1)$
- Or what is the probability that a person is really sick if their st is negative (**false negative**) ? $\mathbb{P}(S = 1|T = 0)$

Before authorizing the marketing of this test, what should be checked?

- A rare disease if affection 0.1% of the population:
  - $\mathbb{P}(S=1) = 0.001$, with $S=1$ the patient is ill.
  - $\mathbb{P}(S=0) = 1 - \mathbb{P}(S=0) = 0.999 = 99,9\%$, with $S=0$ the patient is not ill.
- When administered to an ill person, the test is reliable with a probability of 0.92:
  - $\mathbb{P}(T=1|S=1) = 0.92 \quad \Leftrightarrow \quad \mathbb{P}(T=0|S=1) = 0.08$
- If a person is not ill, the test is positive with a probability of 0.04:
  - $\mathbb{P}(T=1|S=0) = 0.04 \quad \Leftrightarrow \quad \mathbb{P}(T=1|S=0) = 0.96$
- **We are interested in the likelihood of a person getting sick if they test positive**
  - $\mathbb{P}(S=1|T=1)$: did you detect correctly the ill people ?
  - Or $\mathbb{P}(S=1|T=0)$: did you forget to detect sick people ?

Bayes' theorem: Rare disease detection

- Summary of results highlighted by the laboratory:

$$\mathbb{P}(S = 1) = 0.001 \qquad \mathbb{P}(S = 0) = 0.999$$
$$\mathbb{P}(T = 1|S = 1) = 0.92 \qquad \mathbb{P}(T = 1|S = 0) = 0.04$$
$$\mathbb{P}(T = 0|S = 1) = 0.08 \qquad \mathbb{P}(T = 1|S = 0) = 0.96$$

- What is the probability that a person will be sick if the test is positive ?

$$\mathbb{P}(S = 1|T = 1) =?$$

- There is only a 2.25% chance that a person who is positive on the test will actually be sick, the test is absolutely not reliable !

Bayes' theorem: Rare disease detection

- Summary of results highlighted by the laboratory:

$$
\begin{aligned}
\mathbb{P}(S=1) &= 0.001 & \mathbb{P}(S=0) &= 0.999 \\
\mathbb{P}(T=1|S=1) &= 0.92 & \mathbb{P}(T=1|S=0) &= 0.04 \\
\mathbb{P}(T=0|S=1) &= 0.08 & \mathbb{P}(T=1|S=0) &= 0.96
\end{aligned}
$$

- What is the probability that a person will be sick if the test is positive ?

$$
\begin{aligned}
\mathbb{P}(S=1|T=1) &= \frac{\mathbb{P}(S=1)\mathbb{P}(T=1|S=1)}{\mathbb{P}(T=1)} \\
&= \frac{\mathbb{P}(S=1)\mathbb{P}(T=1|S=1)}{\mathbb{P}(T=1,S=0)+\mathbb{P}(T=1,S=1)} \\
&= \frac{\mathbb{P}(S=1)\mathbb{P}(T=1|S=1)}{\mathbb{P}(T=1|S=0)\mathbb{P}(S=0)+\mathbb{P}(T=1|S=1)\mathbb{P}(S=1)} \\
&= \frac{0.001 * 0.92}{0.04 * 0.999 + 0.92 * 0.001} = 0.0225
\end{aligned}
$$

- There is only a 2.25% chance that a person who is positive on the test will actually be sick, the test is absolutely not reliable !

Supervised approach

Observe a dataset $\mathcal{D}$ which can be split as:

- A training dataset $\mathcal{D}_{train}$ composed of labeled observations with d features $\{X_i = (x_{i1}, ..., x_{id})\}_{i\in[N]} \subset \mathbb{R}^d$ and labels $\{y_i\}_{i\in[N]} \subset [C]$ assigning one class c out of C.
- A test dataset $\mathcal{D}_{test}$ only composed of features.
- Classifiers aim at inferring labels of $\mathcal{D}_{test}$ from the knowledge of the annotated dataset $\mathcal{D}_{train}$.
- Feature values can be:
    - Categorical *e.g.* eyes color (blue, green, brown)...
    - Continuous *e.g.* height ...

Naive bayes classifiers: generic method

Take an observation $X_i = (x_{i1}, ..., x_{id})$ from train set $\mathcal{D}_{train}$. To which class c does X_i belong ?

Bayes' approach:

- Estimate probabilities of this sample to belong to any class $c \in [C]$:

$$\begin{aligned}\mathbb{P}(Y = c|X_i) &= \frac{\mathbb{P}(Y = c)\mathbb{P}(X_i|Y = c)}{\mathbb{P}(X_i)} \qquad \text{(Bayes' rule)}\\ &\propto \mathbb{P}(Y = c)\mathbb{P}(X_i|Y = c)\\ &= \mathbb{P}(Y = c)\underbrace{\mathbb{P}(x_1 = x_{i1}, ..., x_d = x_{id}|Y = c)}_{(\star)}\end{aligned}$$

- **Main difficulty**: How to learn the joint probability over features in $(\star)$?

- **Main assumption: All feature components are independent conditionally to Y**

Naive bayes classifiers: generic method

Take an observation $X_i = (x_{i1}, ..., x_{id})$ from train set $\mathcal{D}_{train}$. To which class c does X_i belong ?

Bayes' approach:

- Estimate probabilities of this sample to belong to any class $c \in [C]$:

$$\underbrace{\mathbb{P}(Y = c|X_i)}_{\text{Posterior}} \propto \underbrace{\mathbb{P}(Y = c)}_{\text{Prior}} \underbrace{\prod_{j \in [d]} \mathbb{P}(x_j = x_{ij}|Y = c)}_{\text{Likelihood}}$$

- Prior and Likelihood are estimated over $\mathcal{D}_{train}$, but how? **based on which assumptions ?**
- Prior: Multinomial or Bernouilli distribution depending on multi-class/binary.
- Assumptions on conditional probability $\mathbb{P}(x_j = x_{ij}|Y = c)$ over feature components x_j, for instance:
    - Binary: Bernouilli distribution
    - Categorical: Multinomial distribution
    - Continuous: Gaussian distribution.

Naive bayes classifiers: generic method

Take an observation $X_i = (x_{i1}, ..., x_{id})$ from train set $\mathcal{D}_{train}$. To which class c does X_i belong ?

Bayes' approach:

- Estimate probabilities of this sample to belong to any class $c \in [C]$:

$$\underbrace{\mathbb{P}(Y = c|X_i)}_{\text{Posterior}} \propto \underbrace{\mathbb{P}(Y = c)}_{\text{Estimated Prior}} \underbrace{\prod_{j \in [d]} \mathbb{P}(x_j = x_{ij}|Y = c)}_{\text{Estimated Likelihood}}$$

- **Maximum A Posteriori (MAP) estimation**: Assign to X_i the class $c^\star \in [C]$ which maximizes the posterior probability:

$$X_i \sim y_i = c^\star \leftarrow arg \max_{c \in [C]} \mathbb{P}(Y = c|X_i)$$

Learn a Naive Bayes classifier

We have a dataset of $N = 7$ observations with features $X_i =$(height, weight, foot size, eyes color) of the i-th individual which are used to classify this person as male or female (*e.g.* $y_i \in \{0, 1\} = \{male, female\}$).

Height (cm)	Weight (kg)	Foot Size(cm)	Eyes color	Sex
182	81	31	Blue	male
180	86	27	Brown	male
170	77	31	Blue	male
152	46	16	Blue	female
167	68	20	Blue	female
165	58	17	Blue	female
175	67	22	Brown	female

Naive Bayes classifiers: example

Learn a Naive Bayes classifier:

Height (cm)	Weight (kg)	Foot Size(cm)	Eyes color	Sex
182	81	31	Blue	male
180	86	27	Brown	male
170	77	31	Blue	male
152	46	16	Blue	female
167	68	20	Blue	female
165	58	17	Blue	female
175	67	22	Brown	female

- 1) Estimate the prior distribution $\mathbb{P}(Y)$. How ?

Naive Bayes classifiers: example

Learn a Naive Bayes classifier:

Height (cm)	Weight (kg)	Foot Size(cm)	Eyes color	Sex
182	81	31	Blue	male
180	86	27	Brown	male
170	77	31	Blue	male
152	46	16	Blue	female
167	68	20	Blue	female
165	58	17	Blue	female
175	67	22	Brown	female

- 1) Estimate the prior distribution $\mathbb{P}(Y)$. How ?

  Count occurrences of each label within the dataset:

$$\mathbb{P}(Y = male) = \frac{3}{7}$$

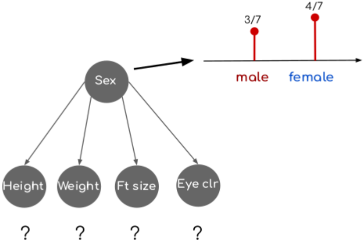
$$\mathbb{P}(Y = female) = 1 - \mathbb{P}(Y = male) = \frac{4}{7}$$

Naive Bayes classifiers: example

Learn a Naive Bayes classifier:

Height (cm)	Weight (kg)	Foot Size(cm)	Eyes color	Sex
182	81	31	Blue	male
180	86	27	Brown	male
170	77	31	Blue	male
152	46	16	Blue	female
167	68	20	Blue	female
165	58	17	Blue	female
175	67	22	Brown	female

- 1) Estimate the prior distribution $\mathbb{P}(Y)$: Count occurrences of each label within the dataset.
- 2) Estimate conditional probabilities $\mathbb{P}(x_j = x_{ij} | Y = c)$, How ?

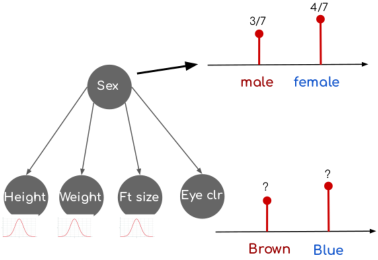


Naive Bayes classifiers: example

Learn a Naive Bayes classifier:

Height (cm)	Weight (kg)	Foot Size(cm)	Eyes color	Sex
182	81	31	Blue	male
180	86	27	Brown	male
170	77	31	Blue	male
152	46	16	Blue	female
167	68	20	Blue	female
165	58	17	Blue	female
175	67	22	Brown	female

- 1) Estimate the prior distribution $\mathbb{P}(Y)$: Count occurrences of each label within the dataset.
- 2) Estimate conditional probabilities $\mathbb{P}(x_j = x_{ij} | Y = c)$, How ?

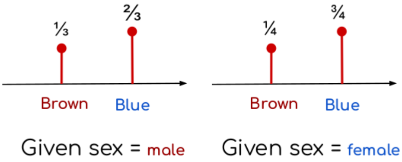


Learn a Naive Bayes classifier:

Height (cm)	Weight (kg)	Foot Size(cm)	Eyes color	Sex
182	81	31	Blue	male
180	86	27	Brown	male
170	77	31	Blue	male
152	46	16	Blue	female
167	68	20	Blue	female
165	58	17	Blue	female
175	67	22	Brown	female

- 1) Estimate the prior distribution $\mathbb{P}(Y)$: Count occurrences of each label within the dataset.
- 2) Estimate conditional probabilities $\mathbb{P}(x_j = x_{ij} | Y = c)$, depending on chosen assumption:

$\mathbb{P}(\text{Eyes color} | Y = c) \sim \mathcal{B}(p_c)$



Naive Bayes classifiers: example

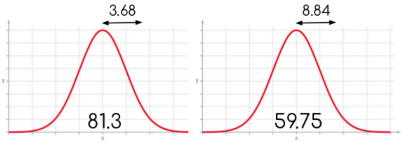
Learn a Naive Bayes classifier:

Height (cm)	Weight (kg)	Foot Size(cm)	Eyes color	Sex
182	81	31	Blue	male
180	86	27	Brown	male
170	77	31	Blue	male
152	46	16	Blue	female
167	68	20	Blue	female
165	58	17	Blue	female
175	67	22	Brown	female

- 1) Estimate the prior distribution $\mathbb{P}(Y)$: Count occurrences of each label within the dataset.

- 2) Estimate conditional probabilities $\mathbb{P}(x_j = x_{ij} | Y = c)$, depending on chosen assumption:

$$\mathbb{P}(\text{Height}|Y = c) \sim \mathcal{N}(\mu_c, \sigma_c)$$



Given sex = male   Given sex = female

Naive Bayes classifiers: example

Learn a Naive Bayes classifier:

Height (cm)	Weight (kg)	Foot Size(cm)	Eyes color	Sex
182	81	31	Blue	male
180	86	27	Brown	male
170	77	31	Blue	male
152	46	16	Blue	female
167	68	20	Blue	female
165	58	17	Blue	female
175	67	22	Brown	female

- 1) Estimate the prior distribution $\mathbb{P}(Y)$: Count occurrences of each label within the dataset.
- 2) Estimate conditional probabilities $\mathbb{P}(x_j = x_{ij} | Y = c)$, depending on chosen assumption. For all feature components.
- 3) Compute posterior probabilities $\mathbb{P}(Y = c | X_i)$ for all c.
- 4) Assign to X_i the class $c^\star \in [C]$ such that:

$$X_i \sim y_i = c^\star \leftarrow arg \max_{c \in [C]} \mathbb{P}(Y = c | X_i)$$

Naive bayes classifiers: Testing phase

Take an observation $X_i = (x_{i1}, ..., x_{id})$ from **the test set** $\mathcal{D}_{test}$. To which class c should X_i belong ?

Bayes' approach:

- Knowing estimates of each posterior probability $\mathbb{P}(Y = c|X \in \mathcal{D}_{train})$
- Compute likelihood of the observation $\mathbb{P}(X_i|Y = c)$ for all c, where parameters of theses distributions are considered known after the training phase.
- Same as training phase (MAP): assign to X_i the class $c^\star \in [C]$ such that:

$$X_i \sim \hat{y}_i = c^\star \leftarrow arg \max_{c \in [C]} \mathbb{P}(Y = c|X_i)$$

Naive bayes classifiers: Testing phase

Take an observation $X_i = (x_{i1}, ..., x_{id})$ from **the test set** $\mathcal{D}_{test}$. To which class c should X_i belong ?

Zero conditional probability problem

- If a feature component value is not contained in the training set and is not supported by the estimated conditional distribution of this feature (*e.g.* green eyes in the previous example.)

$$\mathbb{P}(X_i|Y = c) = \mathbb{P}(Y = c) \prod_{j \in [d]} \mathbb{P}(x_j = x_{ij}|Y = c) = 0$$

- To solve the problem, the probability is estimated using (Laplace smoothing)

$$\mathbb{P}(x_j = x_{ij}|Y = c) = \frac{N_{j,c}(x_{ij}) + \lambda}{N_c + \lambda N}$$

  - $N_{j,c}(x_{ij})$: number of training example for which $x_j = x_{ij}$ and $Y = c$.
  - N_c: number of training examples such that $Y = c$.
  - N: number of observations in the training set.
  - $\lambda > 0$: smoothing parameter.

Naive-Bayes classifiers: conclusion

- **Independence assumption: All feature components are independent conditionally to Y:**
  - For many real work tasks $\mathbb{P}(x_1, ..., x_d|Y) \neq \mathbb{P}(x_1|Y)...\mathbb{P}(x_d|Y)$!
    e.g. height and foot size are somehow correlated in the sex classification task.
  - Each distribution can be independently estimated as a one dimensional distribution. This in turn helps to alleviate problems stemming from the curse of dimensionality.

- The different Naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of $\mathbb{P}(x_i|y)$

  - **Multinomial Naive Bayes**:
    Used e.g. for document classification, like whether a document belongs to the category of sports, politics, technology etc. The features/predictors used by the classifier are the frequency of the words present in the document.
  - **Bernouilli Naive Bayes**:
    Similar to the Multinomial one but the features are boolean variables. the parameters that we use o predict the class variable take up only values yes or no, e.g. if a word occurs in the text or not.
  - **Gaussian Naive Bayes**:
    When predictors take up a continuous value and are not discrete, we assume that these values are sampled from a gaussian distribution.

- Nevertheless, Naive Bayes classifier works well in many real-world situations:

  e.g. document classification and spam filtering.
- Also Naive Bayes can perform well on classification, it is known to be a bad estimator, so the posterior probability outputs are not that reliable.
- Naive Bayes is not naturally suitable for regression:
  - These models can be considered a way of fitting a probability model that optimizes the joint likelihood $\mathbb{P}(X, Y)$.
  - Some work try to use Naive Bayes for regression using kernel density estimators https://www.cs.waikato.ac.nz/ eibe/pubs/nbr.pdf.

Naive-Bayes classifiers: pros and cons

Pros

- Computationally fast: $0(n_{features} \times n_{samples})$
- Simple to implement
- Works well with small datasets
- Works well with high dimensions
- Perform well even if the independence assumption is not perfectly met. In many cases, the approximation is enough to build a good classifier.

Cons

- Require to **remove correlated features** because they are voted twice in the model and it can lead to **over inflating importance**.
- Issue if a categorical variable has a category in test set which was not observed in the training set $\rightarrow$ Zero conditional probability problem.
    - To solve this, we can use smoothing techniques such as the Laplace smoothing (default in scikit-learn).