CSE517A Machine Learning **Spring 2020**

Lecture 5: Naive Bayes Classifier

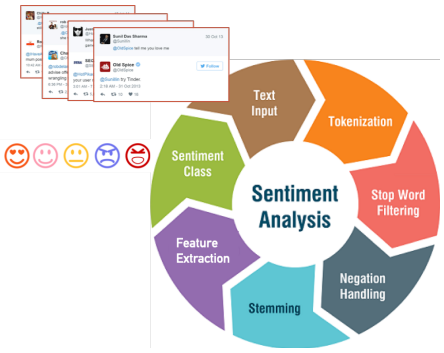
Instructor: Marion Neumann *Scribe: Jingyu Xin*

Reading: FCML 5.2.1 (Bayes Classifier, Naive Bayes, Classifying Text, and Smoothing)

Application

Sentiment analysis aims at discovering people's opinions, emotions, feelings about a subject matter, product, or service from text.

Take some time to answer the following warm-up questions: (1) Is this a regression or classification problem? (2) What are the features and how would you represent them? (3) What is the prediction task? (4) How well do you think a linear model will perform?



1 Introduction

Thought: Can we model $p(y \mid \mathbf{x})$ without model assumptions, e.g. Gaussian/Bernoulli distribution etc.?

Idea: Estimate $p(y \mid \mathbf{x})$ from the data *directly*, then use the Bayes classifier.

Let y be discrete,

$$p(y \mid \mathbf{x}) = \frac{\sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x} \cap y_i = y)}{\sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x})} \tag{1}$$

where the numerator counts all examples with input $\mathbf{x}$ that have label y and the denominator counts all examples with input $\mathbf{x}$.

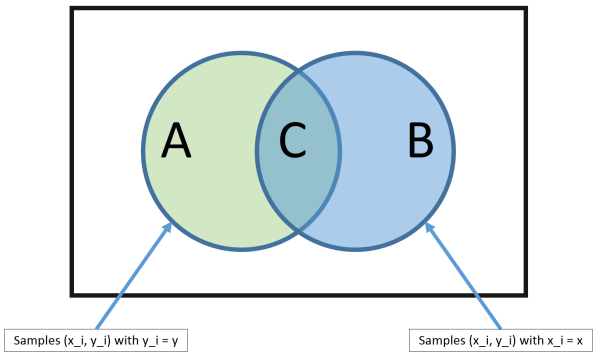


Figure 1: Illustration of estimating $\hat{p}(y \mid \mathbf{x})$

From Figure ??, it is clear that, using the MLE method, we can estimate $\hat{p}(y \mid \mathbf{x})$ as

$$\hat{p}(y \mid \mathbf{x}) = \frac{|C|}{|B|} \tag{2}$$

But there is a big **problem** with this method:
The MLE estimate is only good if there are many training vectors with the **same identical** features as $\mathbf{x}$.

This **never** happens for high-dimensional or continuous feature spaces.

Solution: Bayes Rule.

$$p(y \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y)\, p(y)}{p(\mathbf{x})} \tag{3}$$

Let's estimate $p(\mathbf{x} \mid y)$ and $p(y)$ instead!

2 Naive Bayes

We have a discrete label space C that can either be *binary* $\{+1, -1\}$ or *multi-class* $\{1, ..., K\}$. The feature space $\mathcal{X}$ may be *categorical* (binary or with more than two categories), *multi-nomial*, or *continuous*.

2.1 Estimate $p(y)$

Estimating $p(y)$ is easy. If y takes on discrete binary values, for example, this just becomes coin tossing.

Let $\hat{\pi}$ be our estimator for $p(y)$ derived via counting similar as illustrated for $p(y \mid \mathbf{x})$ above.

Then, for *binary classification* we have:

$$\hat{\pi}_{y=+1} = p(y = +1)$$
$$\hat{\pi}_{y=-1} = p(y = -1)$$

and for *multi-class classification* we get:

$$\hat{\pi}_c = p(y = c) = \frac{\sum_{i=1}^{n} I(y_i = c)}{n} \qquad \text{for all } c \in \{1, ..., K\}.$$

2.2 Estimate $p(\mathbf{x} \mid y)$

Estimating $p(\mathbf{x} \mid y)$ is not easy.
Question: Why?

So, we have to make an assumption, the so-called *Naive Bayes* assumption.

Naive Bayes Assumption: The features are **independent given the label**,

$$p(\mathbf{x} \mid y) = \prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha \mid y) \tag{4}$$

where $[\mathbf{x}]_\alpha$ is the value for feature α.

Assume Equation (4) holds, then the **Naive Bayes classifier** can be derived from the **Bayes classifier** as:

$$\begin{aligned} y = h(\mathbf{x}) &= \arg\max_y p(y \mid \mathbf{x}) \\ &= \arg\max_y \frac{p(\mathbf{x} \mid y) p(y)}{p(\mathbf{x})} \\ &= \arg\max_y p(\mathbf{x} \mid y)\, p(y) \\ &= \arg\max_y \prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha \mid y)\, p(y) \end{aligned} \tag{5}$$

Estimating $\log p([\mathbf{x}]_\alpha \mid y)$ is easy as we only need to consider one dimension. And estimating $p(y)$ is not affected by the assumption.

Case 1: Categorical Features

Features: $[\mathbf{x}]_\alpha \in \{1, 2, ..., K_\alpha\}$ (no ordering, K_α = number of categories for feature α).

Model: We use the categorical distribution (also sometimes referred to *multinoulli* distribution):

$$p([\mathbf{x}]_\alpha \mid y = c) = \prod_{j=1}^{K_\alpha} [\boldsymbol{\theta}_{\alpha c}]_j^{I([\mathbf{x}]_\alpha = j)} \tag{6}$$

with parameters $[\boldsymbol{\theta}_{\alpha c}]_j = p([\mathbf{x}]_\alpha = j \mid y = c)$.

$\boldsymbol{\theta}_{\alpha c}$ is the vector of category probabilities of input feature α, given that the class label is c. That is, the j-th entry in this probability vector $[\boldsymbol{\theta}_{\alpha c}]_j$ is the probability of feature α having the value j, given that the label is c. Note that these probabilities sum to one:

$$\sum_{j=1}^{K_\alpha} [\boldsymbol{\theta}_{\alpha c}]_j = 1.$$

Parameter Estimation:

$$[\hat{\boldsymbol{\theta}}_{\alpha c}]_j = \frac{\sum_{i=1}^n I(y_i = c) I([\mathbf{x}_i]_\alpha = j) + l}{\sum_{i=1}^n I(y_i = c) + l K_\alpha} \tag{7}$$

where l is a *smoothing* parameter.

Training the Naive Bayes classifier corresponds to estimating θ_{jc} for all j and c and storing them in the respective conditional probability tables (CPT). Also note that by setting $l = 0$, we get an MLE estimator; $l > 0$ leads to MAP. Setting $l = 1$ is referred to as *Laplace smoothing*.

Prediction:

$$\begin{aligned} y &= \arg\max_c p(y = c \mid \mathbf{x}) \\ &= \arg\max_c \hat{\pi}_c \prod_{\alpha=1}^{d} [\hat{\boldsymbol{\theta}}_{\alpha c}]_j \end{aligned} \tag{8}$$

Exercise 2.1. Should we play tennis or not?

Consider the test input $\mathbf{x}_* = [Overcast, Hot, High, Strong]$, work out the CPTs and $\hat{\pi}$, then predict y_* (y_* =yes or y_* =no) for the following data:

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Case 2: Multinomial Features

Features: $[\mathbf{x}]_\alpha \in \{0, 1, 2, ..., m\}$ (count features).
Each feature $[\mathbf{x}]_\alpha$ are counts of occurrences in a sequence of length $m = \sum_{\alpha=1}^{d} [\mathbf{x}]_\alpha$. An example of this is the representation of text, where the sequence is the document and the features are counts of each specific word α in the document of length m. The feature dimensionality d is the size of the vocabulary.

Model: Use the multinomial distribution (cf. FCML 2.3.3):

$$p(\mathbf{x} \mid m, y = c) = \frac{m!}{[\mathbf{x}]_1![\mathbf{x}]_2!...[\mathbf{x}]_d!} \prod_{\alpha=1}^{d} \theta_{\alpha c}^{[\mathbf{x}]_\alpha} \tag{9}$$

where $\theta_{\alpha c}$ is the probability of selecting $[\mathbf{x}]_\alpha$ and $\sum_{\alpha=1}^{d} \theta_{\alpha c} = 1$.

For example, we can use this to generate a spam email, $\mathbf{x}$ of class $y =$ spam by by picking m words independently at random from the vocabulary of d word using $p(\mathbf{x} \mid y = spam)$.

Parameter Estimation:

$$\hat{\theta}_{\alpha c} = \frac{\sum_{i=1}^{n} I(y_i = c)[\mathbf{x}_i]_\alpha + l}{\sum_{i=1}^{n} I(y_i = c) \sum_{\beta=1}^{d} [\mathbf{x}_i]_\beta + ld} \tag{10}$$

where the numerator sums up all counts for feature α given the class label c and the denominator sums up all counts of all features given the class label. E.g.:

$$\frac{\text{\# of times word } \alpha \text{ appears in all spam emails}}{\text{length of all spam emails}}$$

Note that $\sum_{\beta=1}^{d} [\mathbf{x}_i]_\beta$ corresponds to m_i and l is a *smoothing* parameter.

Prediction:

$$\begin{aligned} y &= \arg\max_c p(y = c \mid \mathbf{x}) \\ &= \arg\max_c \hat{\pi}_c \prod_{\alpha=1}^{d} \hat{\theta}_{\alpha c}^{[\mathbf{x}]_\alpha} \end{aligned} \tag{11}$$

Exercise 2.2. Consider the following **sentiment analysis** data:

"this book is awesome"	*positive*
"harry potter books suck"	*negative*
"these pretzles are making me thirsty"	*negative*
"they choppin my fingers off Ira"	*negative*
"supreme beings of leisure rock"	*positive*
"cheeto jesus is a tyrant"	*negative*

(a) Train a multinomial Naive Bayes classifier for binary sentiment prediction ($y \in \{positive, negative\}$). HINT: Perform some text preprocessing, like stopword filtering.

(b) Using your trained naive Bayes model, predict the sentiment for the following test input x^* = "just had my first cheeto ever it was awesome".

(c) Add a smoothing term of $l = 1$ to the probability of observing each word. Recalculate your Naive Bayes prediction.

Case 3: Continuous Features (Gaussian Naive Bayes)

Features: $[\mathbf{x}]_\alpha \in \mathbb{R}$ (the feature takes on a real value).

Model: Use Gaussian distribution (cf. FCML 2.5.3):

$$p([\mathbf{x}]_\alpha \mid y = c) = \mathcal{N}(\mu_{\alpha c}, \sigma^2_{\alpha c}) = \frac{1}{\sqrt{2\pi\sigma^2_{\alpha c}}} e^{-\frac{1}{2}\left(\frac{[\mathbf{x}]_\alpha - \mu_{\alpha c}}{\sigma_{\alpha c}}\right)^2} \tag{12}$$

Note that the model specified above is based on our assumption about the data - that each feature α comes from a class-conditional Gaussian distribution. This model choice leads to Gaussian Naive Bayes (GNB). Other specification could be used as well.

Parameter Estimation:

$$\hat{\mu}_{\alpha c} \leftarrow \frac{1}{n_c} \sum_{i=1}^{n} I(y_i = c)[\mathbf{x}_i]_\alpha \tag{13}$$

$$\hat{\sigma}^2_{\alpha c} \leftarrow \frac{1}{n_c} \sum_{i=1}^{n} I(y_i = c) \left([\mathbf{x}_i]_\alpha - \mu_{\alpha c}\right)^2 \tag{14}$$

where $n_c = \sum_{i=1}^{n} I(y_i = c)$.

Exercise 2.3. Consider the following data:

	Gender	Height	Weight	Foot_Size
0	male	6.00	180	12
1	male	5.92	190	11
2	male	5.58	170	12
3	male	5.92	165	10
4	female	5.00	100	6
5	female	5.50	150	8
6	female	5.42	130	7
7	female	5.75	150	9

(a) Train a Gaussian Naive Bayes classifier for a binary gender prediction task ($y \in \{male, female\}$).

(b) Using your trained naive Bayes model, predict the gender for the following test input $x^* = [6, 130, 8]$.

(c) Why do we not need to add smoothing terms for continuous features?

2.3 Naive Bayes is a Linear Classifier

If the likelihood factors $p([\mathbf{x}]_\alpha \mid y)$ are from the *exponential family* (e.g. Gaussian, Bernoulli, Dirichlet, beta, multinomial (with fixed number of trials)...), Naive Bayes is a linear classifier. In general, however, this is not true.

(1) Suppose $y_i \in \{-1, +1\}$ and features are multinomial. We can show that

$$\begin{aligned} h(\mathbf{x}) &= \arg\max_y p(y) \prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha \mid y) \\ &= sign(\mathbf{w}^\top \mathbf{x} + b) \end{aligned} \tag{15}$$

That is

$$h(\mathbf{x}) = +1 \iff \mathbf{w}^\top \mathbf{x} + b > 0 \tag{16}$$

(2) In the case of continuous features (GNB) and $y_i \in \{-1, +1\}$, let $\sigma_{\alpha,c=-1} = \sigma_{\alpha,c=+1}$ (standard deviation does not depend on label), we can show that

$$p(y \mid \mathbf{x}) = \frac{1}{1 + e^{-y(\mathbf{w}^\top \mathbf{x} + b)}} \tag{17}$$

Note: This model is also known as logistic regression. NB and LR produce *asymptotically* the same model. But NB and LR are different learning algorithms since the parameters are estimated differently.

2.4 [optional] "True" Bayesian Naive Bayes

Note that even though Naive Bayes uses Bayes rule, it is **not** a Bayesian classifier. We can make it Bayesian by incorporating a *prior distribution over our parameters*. As an example we will look at *multi-class* classification with *binary* features here.

Similar to the Bayesian approach to coin flipping, we can incorporate a prior distribution over $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ and utilize the posterior distribution for predictions to achieve "true" Bayesian Naive Bayes. Common choices for the prior distributions are the Dirichlet distribution for the class probability (multivariate generalization of the beta distribution) and the factored beta distribution for $\boldsymbol{\theta}$:

$$\begin{aligned} p(\boldsymbol{\pi}) &= Dir(\boldsymbol{\alpha}) \\ p(\boldsymbol{\theta}) &= \prod_{\alpha} \prod_{c} p(\theta_{\alpha c}), \text{ where} \\ p(\theta_{\alpha c}) &= Beta(\beta_0, \beta_1) \end{aligned}$$

where α, β_0, β_1 are the *hyperparameters*. If we set those to 1, we get Laplace smoothing which is equivalent to MAP estimation. With our Bayesian framework, we can now use the posterior means $\bar{\boldsymbol{\pi}}$ and $\bar{\theta}_{\alpha c} \, \forall \alpha, c$ to get estimates for $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$. Those posterior means can again be computed via counting and their derivation is left as an exercise for the interested reader. Note that for categorical features $\boldsymbol{\theta}_{\alpha c}$ would be a vector again and we would need the Dirichlet distribution instead of the beta distribution.

Prediction:

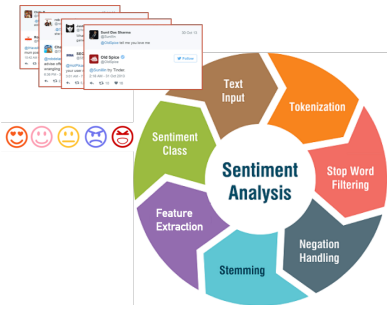
$$\begin{aligned} p(y = c \mid \mathbf{x}, D) &\propto \int p(y = c \mid \pi)\, p(\pi \mid D)\, \mathrm{d}\pi \prod_{\alpha=1}^{d} \int p([\mathbf{x}]_\alpha \mid y = c, \theta_{\alpha c})\, p(\theta_{\alpha c} \mid D)\, \mathrm{d}\theta_{\alpha c} \\ &\propto \bar{\pi}_c \prod_{\alpha=1}^{d} (\bar{\theta}_{\alpha c})^{I([\mathbf{x}]_\alpha = j)} (1 - \bar{\theta}_{\alpha c})^{I([\mathbf{x}]_\alpha \neq j)} \end{aligned} \tag{18}$$

2.5 Mixed and Missing Features

Naive Bayes can handle features of *mixed types* and even *missing features* very naturally. To see the former, recall that the Naive Bayes classifier is given by $h(\mathbf{x}) = \arg\max_y \prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha \mid y)\, p(y)$ (cf. Eq. (5)) and that we can estimate $p([\mathbf{x}]_\alpha \mid y)$ according to each feature type independently.

Exercise 2.4. To see how the Naive Bayes classifier can deal with unobserved (missing) features we will look at an example: **Suzie's date** (in-class exercise; also available on Canvas).

Application



Reconsider our sentiment analysis problem.

If you were to use Naive Bayes as your sentiment predictor, what feature representation would make sense: *categorical*, *multinomial*, *continuous*, or mixed? (There is no correct/incorrect answer, so be creative!)

Is your Naive Bayes sentiment predictor a linear classifier?

We will implement a Naive Bayes classifier in our second implementation project.

2.6 Summary

- MLE overfits and does not work for small n (number of samples)
- use smoothing $\rightarrow$ MAP solution
- Naive Bayes is not a Bayesian method
- Naive Bayes can deal with mixed and missing features (this is not the case for most other ML methods!)

Exercise 2.5. Using *your own words*, recap each of the above summary points in 1-2 sentences.[a]

[a] *Yes,* ***say it out loud*** *or* ***write it down****, it'll help you retain the knowledge!*

3 Discussion: Generative v.s. Discriminative Learning

All models use $p(y \mid \mathbf{x})$ for predictions, but

- *generative* models estimate $p(\mathbf{x} \mid y)\, p(y)$
- *discriminative* models estimate $p(y \mid \mathbf{x})$ directly

Here we give some comparisons between generative and discriminative learning.

(1) Training efficiency: generative models are usually faster to train

(2) Predicition quality: discriminative models are usually more accurate and give better probability estimates

(3) Generating data: generative models can be used to generate data (e.g. spam email)

(4) Assumptions: generative models make strong independence assumptions which are often not valid; discriminative models make stronger model assumptions