NAIVE BAYES CLASSIFIERS

Recall

Two random variables are *stochastically independent*, or *independent* for short, if their joint distribution factorizes:

$$P(x, y) = P(x)P(y) \qquad \text{or} \qquad p(x, y) = p(x)p(y)$$

Dependent means *not independent*.

Intuitively

X and Y are dependent if knowing the outcome of X provides any information about the outcome of Y.

More precisely:

- If someone draws (X, Y) simultaneously, and only discloses $X = x$ to you, does that change your mind about the distribution of Y? (If so: Dependence.)
- Once X is given, the conditional distribution of Y is $P(Y|X = x)$.
- If that is still $P(Y = y)$, as before X was drawn, the two are independent. If $P(Y|X = x) \neq P(Y)$, they are dependent.

A few remarks

- Joint distributions of dependent variables can become very complicated. Dealing with joint distributions of many variables is one of the hardest problems in statistics and probability.
- The math almost always becomes easier if we assume variables are independent.
- On the other hand, assuming independence means we neglect all interactions between the effects represented by the variables.
- When we design probability models, there is usually a trade-off between simplicity (e.g. assuming everything is independent) and accuracy (trying to model all interactions precisely).

Simplest form

- Random variables $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$, where $\mathbf{X}, \mathbf{Y}$ are finite sets.
- Each possible value of X and Y has positive probability.

Then

$$P(X = x, Y = y) = P(y|x)P(x) = P(x|y)P(y)$$

and we obtain

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} = \frac{P(x|y)P(y)}{\sum_{y \in \mathcal{Y}} P(x|y)P(y)}$$

It is customary to name the components,

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

In terms of densities

For continuous sets $\mathbf{X}$ and $\mathbf{Y}$,

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int_{\mathbf{Y}} p(x|y)dy}$$

Classification

We define a classifier as

$$f(\mathbf{x}) := \arg\max_{y \in [K]} P(y|\mathbf{x})$$

where $\mathbf{Y} = [K]$ and $\mathbf{X}$ = sample space of data variable.

With the Bayes equation, we obtain

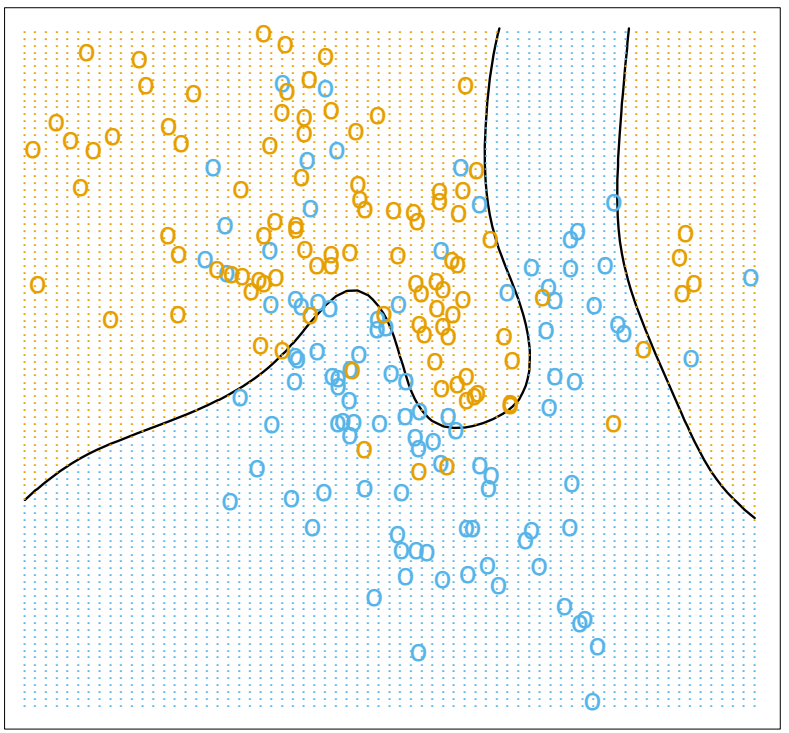
$$f(\mathbf{x}) = \arg\max_{y} \frac{p(\mathbf{x}|y)P(y)}{p(\mathbf{x})} = \arg\max_{y} p(\mathbf{x}|y)P(y)$$

If the class-conditional distribution is continuous, we use

$$f(\mathbf{x}) = \arg\max_{y} p(\mathbf{x}|y)P(y)$$

Optimal classifier

- In the risk framework, the best possible classifier is the one which minimizes the risk.
- Which classifier is optimal depends on the chosen cost function.



Zero-one loss

Under zero-one loss, the classifier which minimizes the risk is the classifier

$$f(\mathbf{x}) = \arg\max_y P(x|y)P(y)$$

from the previous slide. When computed from the *true* distribution of (X, Y), this classifier is called the **Bayes-optimal classifier** (or **Bayes classifier** for short).

BAYES-OPTIMAL CLASSIFIER

Suppose for simplicity we have to classes labeled "1" and "2", so $y \in \{1, 2\}$.

$$f(\mathbf{x}) = \arg\max_y p(x|y)P(y)$$

What do the terms mean?

- $P(y)$ = probability to observe class $Y = y$ if we draw (X, Y) from $p(x, y)$ and discard X.
- Approximately, this is the probability that a training data point is labeled y if we draw it uniformly from a very large training set (without looking at x).
- If both classes are equally probable (in terms of training data: equally large), then $P(y) = \frac{1}{2}$.
- $P(y|x)$: Fix a point x is space. What is the probability that a data point at this location belongs to class y?
- This is a number strictly between 0 and 1 if the classes "overlap" in space.

If classes are assumed equally large

$$f(\mathbf{x}) = \arg\max_y p(x|y)P(y) = \arg\max_y p(x|y)\,\tfrac{1}{2} = \arg\max_y p(x|y)$$

That means: The Bayes-optimal classifier is the one that assigns a point at location x to the class whose probability at x is larger, e.g. to class 1 if $P(1|x) \geq P(2|x)$.

EXAMPLE: SPAM FILTERING

Representing emails

- $\mathbf{Y} = \{$ spam, email $\}$
- $\mathbf{X} = \mathbb{R}^d$
- Each axis is labeled by one possible word.
- d = number of distinct words in vocabulary
- x_j = number of occurrences of word j in email represented by $\mathbf{x}$

For example, if axis j represents the term "the", $x_j = 3$ means that "the" occurs three times in the email $\mathbf{x}$. This representation is called a **vector space model of text**.

Example dimensions

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

With Bayes equation

$$f(\mathbf{x}) = \underset{y \in \{\text{spam,email}\}}{\operatorname{argmax}} P(y|\mathbf{x}) = \underset{y \in \{\text{spam,email}\}}{\operatorname{argmax}} p(\mathbf{x}|y)P(y)$$

Simplifying assumption

The classifier is called a **naive Bayes** classifier if it assumes

$$p(\mathbf{x}|y) = \prod_{j=1}^{d} p_j(x_j|y) \quad \text{for } \mathbf{x} = (x_1, \ldots, x_d) \ ,$$

i.e. if it treats the individual dimensions of $\mathbf{x}$ as conditionally independent given y.

In spam example

- Corresponds to the assumption that the number of occurrences of a word carries information about y.
- Co-occurrences (how often do given combinations of words occur?) is neglected.

Class prior

The distribution $P(y)$ is easy to estimate from training data:

$$P(y) = \frac{\#\text{observations in class } y}{\#\text{observations}}$$

Class-conditional distributions

The class conditionals $p(x|y)$ usually require a modeling assumption. Under a given model:

- Separate the training data into classes.
- Estimate $p(x|y)$ on class y by maximum likelihood.

Class-conditional in the spam example

$P(x|y)$ is a multinomial (= categorical distribution). It is estimated as:

$$P(\text{word } i|y) = \frac{\#\text{ occurrences of word } i \text{ in emails of class } y}{\#\text{ occurrences of word } i \text{ in all emails}}$$