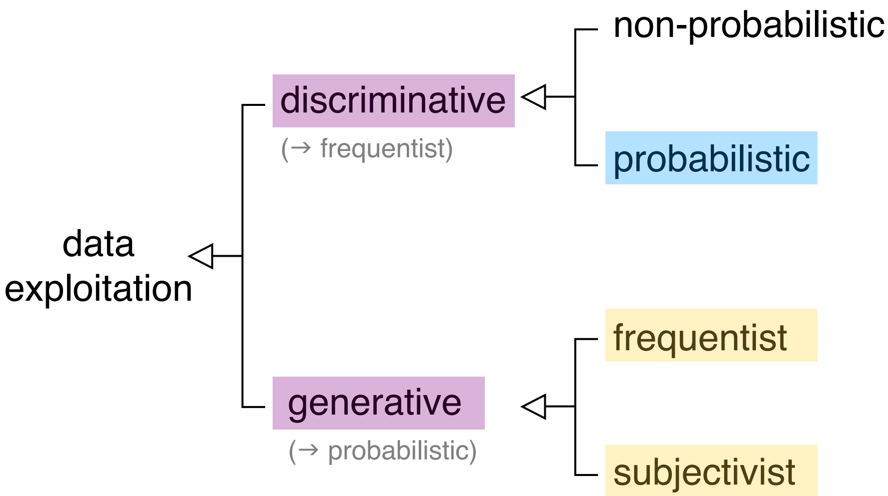
# Chapter ML:VII

## VII. Bayesian Learning

- Approaches to Probability
- Conditional Probability
- Bayes Classifier
- Exploitation of Data
- Frequentist versus Subjectivist

# Frequentist versus Subjectivist

## Logistic Regression versus Naive Bayes [data exploitation examples]




Support vector machine

(1) Linear regression with least square estimates from $D$

(2) Logistic regression via $p()$ with ML estimates from $D$

(3) Logistic regression via $L()$ with ML estimates from $D$

(4) Bayes with prior probability estimates from $D$

(5) Probability model with ML estimate from $D$

(6) Bayes with subjective priors

(4) Bayes with subjective priors

$D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, $D = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_n, c_n)\}$

$D = \{y_1, \ldots, y_n\}$, $D = \{c_1, \ldots, c_n\}$

# Frequentist versus Subjectivist

## Logistic Regression versus Naive Bayes (continued) [data exploitation examples]

(2) $\mathbf{w}_{\text{ML}} = \underset{\mathbf{w} \in \mathbf{R}^{p+1}}{\text{argmax}} \prod_{(\mathbf{x},c) \in D} p(c \mid \mathbf{x}; \mathbf{w}) \rightsquigarrow y(\mathbf{x}) = \sigma(\mathbf{w}_{\text{ML}}^T \mathbf{x}) \rightsquigarrow c_{\mathbf{w}_{\text{ML}}}$ (logistic regression)

(4) $c_{\text{MAP}} = \underset{c \in \{\oplus, \ominus\}}{\text{argmax}} \; p(c \mid \mathbf{x})$

Observation 1. Both approaches maximize $p(D)$ :

- (2), the MLE principle, determines the parameters $\mathbf{w}$ of the logistic model function such that $\prod_D p(c \mid \mathbf{x})$ becomes maximum. Note that a parameter vector $\mathbf{w}$ that maximizes $\prod_D p(c \mid \mathbf{x})$ will also maximize $\prod_D p(\mathbf{x}, c)$, and thus $p(D)$ (under the i.i.d. assumption).

- (4), Naive Bayes, determines for a given $\mathbf{x}$ its most probable class directly. As an application of the Bayesian framework, it chooses $c_{\text{MAP}}$ for each $\mathbf{x}$ and maximizes $p(D)$ by maximizing each factor of $\prod_D p(c \mid \mathbf{x})$. Note that $p(\mathbf{x})$ is constant per factor. Naive Bayes approximates $p(\mathbf{x} \mid c)$ with $\prod_{j=1}^p p(x_j \mid c)$.

# Frequentist versus Subjectivist

## Logistic Regression versus Naive Bayes (continued) [data exploitation examples]

(2) $\mathbf{w}_{\text{ML}} = \underset{\mathbf{w} \in \mathbf{R}^{p+1}}{\operatorname{argmax}} \prod_{(\mathbf{x},c) \in D} p(c \mid \mathbf{x}; \mathbf{w}) \rightsquigarrow y(\mathbf{x}) = \sigma(\mathbf{w}_{\text{ML}}^T \mathbf{x}) \rightsquigarrow c_{\mathbf{w}_{\text{ML}}}$ (logistic regression)

(4) $c_{\text{MAP}} = \underset{c \in \{\oplus, \ominus\}}{\operatorname{argmax}} \dfrac{p(\mathbf{x} \mid c) \cdot p(c)}{p(\mathbf{x})}$ (Bayes)

Observation 1. Both approaches maximize $p(D)$ :

- (2), the MLE principle, determines the parameters $\mathbf{w}$ of the logistic model function such that $\prod_D p(c \mid \mathbf{x})$ becomes maximum. Note that a parameter vector $\mathbf{w}$ that maximizes $\prod_D p(c \mid \mathbf{x})$ will also maximize $\prod_D p(\mathbf{x}, c)$, and thus $p(D)$ (under the i.i.d. assumption).

- (4), Naive Bayes, determines for a given $\mathbf{x}$ its most probable class directly. As an application of the Bayesian framework, it chooses $c_{\text{MAP}}$ for each $\mathbf{x}$ and maximizes $p(D)$ by maximizing each factor of $\prod_D p(c \mid \mathbf{x})$. Note that $p(\mathbf{x})$ is constant per factor. Naive Bayes approximates $p(\mathbf{x} \mid c)$ with $\prod_{j=1}^{p} p(x_j \mid c)$.

# Frequentist versus Subjectivist

## Logistic Regression versus Naive Bayes (continued) [data exploitation examples]

(2) $\mathbf{w}_{\text{ML}} = \underset{\mathbf{w} \in \mathbf{R}^{p+1}}{\operatorname{argmax}} \prod_{(\mathbf{x},c) \in D} p(c \mid \mathbf{x}; \mathbf{w}) \leadsto y(\mathbf{x}) = \sigma(\mathbf{w}_{\text{ML}}^T \mathbf{x}) \leadsto c_{\mathbf{w}_{\text{ML}}}$ (logistic regression)

(4) $c_{\text{MAP}} = \underset{c \in \{\oplus, \ominus\}}{\operatorname{argmax}} \; p(\mathbf{x} \mid c) \cdot p(c)$ (Bayes)

Observation 1. Both approaches maximize $p(D)$ :

- (2), the MLE principle, determines the parameters $\mathbf{w}$ of the logistic model function such that $\prod_D p(c \mid \mathbf{x})$ becomes maximum. Note that a parameter vector $\mathbf{w}$ that maximizes $\prod_D p(c \mid \mathbf{x})$ will also maximize $\prod_D p(\mathbf{x}, c)$, and thus $p(D)$ (under the i.i.d. assumption).

- (4), Naive Bayes, determines for a given $\mathbf{x}$ its most probable class directly. As an application of the Bayesian framework, it chooses $c_{\text{MAP}}$ for each $\mathbf{x}$ and maximizes $p(D)$ by maximizing each factor of $\prod_D p(c \mid \mathbf{x})$. Note that $p(\mathbf{x})$ is constant per factor. Naive Bayes approximates $p(\mathbf{x} \mid c)$ with $\prod_{j=1}^{p} p(x_j \mid c)$.

# Frequentist versus Subjectivist

## Logistic Regression versus Naive Bayes (continued) [data exploitation examples]

(2) $\mathbf{w}_{\text{ML}} = \underset{\mathbf{w} \in \mathbf{R}^{p+1}}{\text{argmax}} \prod_{(\mathbf{x},c) \in D} p(c \mid \mathbf{x}; \mathbf{w}) \rightsquigarrow y(\mathbf{x}) = \sigma(\mathbf{w}_{\text{ML}}^T \mathbf{x}) \rightsquigarrow c_{\mathbf{w}_{\text{ML}}}$ (logistic regression)

(4) $c_{\text{MAP}} = \underset{c \in \{\oplus, \ominus\}}{\text{argmax}} \prod_{j=1}^{p} p(x_j \mid c) \cdot p(c)$ (Naive Bayes)

Observation 1. Both approaches maximize $p(D)$ :

- (2), the MLE principle, determines the parameters $\mathbf{w}$ of the logistic model function such that $\prod_D p(c \mid \mathbf{x})$ becomes maximum. Note that a parameter vector $\mathbf{w}$ that maximizes $\prod_D p(c \mid \mathbf{x})$ will also maximize $\prod_D p(\mathbf{x}, c)$, and thus $p(D)$ (under the i.i.d. assumption).

- (4), Naive Bayes, determines for a given $\mathbf{x}$ its most probable class directly. As an application of the Bayesian framework, it chooses $c_{\text{MAP}}$ for each $\mathbf{x}$ and maximizes $p(D)$ by maximizing each factor of $\prod_D p(c \mid \mathbf{x})$. Note that $p(\mathbf{x})$ is constant per factor. Naive Bayes approximates $p(\mathbf{x} \mid c)$ with $\prod_{j=1}^{p} p(x_j \mid c)$.

# Frequentist versus Subjectivist

## Logistic Regression versus Naive Bayes (continued) [data exploitation examples]

(2) $\mathbf{w}_{\text{ML}} = \underset{\mathbf{w} \in \mathbf{R}^{p+1}}{\text{argmax}} \prod_{(\mathbf{x},c) \in D} p(c \mid \mathbf{x}; \mathbf{w}) \rightsquigarrow y(\mathbf{x}) = \sigma(\mathbf{w}_{\text{ML}}^T \mathbf{x}) \rightsquigarrow c_{\mathbf{w}_{\text{ML}}}$ (logistic regression)

(4) $c_{\text{MAP}} = \underset{c \in \{\oplus, \ominus\}}{\text{argmax}} \prod_{j=1}^{p} p(x_j \mid c) \cdot p(c)$ (Naive Bayes)

Observation 1. Both approaches maximize $p(D)$ :

- (2), the MLE principle, determines the parameters $\mathbf{w}$ of the logistic model function such that $\prod_D p(c \mid \mathbf{x})$ becomes maximum. Note that a parameter vector $\mathbf{w}$ that maximizes $\prod_D p(c \mid \mathbf{x})$ will also maximize $\prod_D p(\mathbf{x}, c)$, and thus $p(D)$ (under the i.i.d. assumption).

- (4), Naive Bayes, determines for a given $\mathbf{x}$ its most probable class directly. As an application of the Bayesian framework, it chooses $c_{\text{MAP}}$ for each $\mathbf{x}$ and maximizes $p(D)$ by maximizing each factor of $\prod_D p(c \mid \mathbf{x})$. Note that $p(\mathbf{x})$ is constant per factor. Naive Bayes approximates $p(\mathbf{x} \mid c)$ with $\prod_{j=1}^{p} p(x_j \mid c)$.

# Frequentist versus Subjectivist

## Logistic Regression versus Naive Bayes (continued) [data exploitation examples]

(2) $\mathbf{w}_{\text{ML}} = \underset{\mathbf{w} \in \mathbf{R}^{p+1}}{\text{argmax}} \prod_{(\mathbf{x},c) \in D} p(c \mid \mathbf{x}; \mathbf{w}) \rightsquigarrow y(\mathbf{x}) = \sigma(\mathbf{w}_{\text{ML}}^T \mathbf{x}) \rightsquigarrow c_{\mathbf{w}_{\text{ML}}}$ (logistic regression)

(4) $c_{\text{MAP}} = \underset{c \in \{\oplus, \ominus\}}{\text{argmax}} \prod_{j=1}^{p} p(x_j \mid c) \cdot p(c)$ (Naive Bayes)

Observation 2 (corollary). Both approaches model the covariate distribution:

- (2), the MLE principle, considers $p(\mathbf{x})$, the distribution of the independent variables $\mathbf{x}$, implicitly via the multiplicity of $\mathbf{x}$ in the data $D$. Recall that $D$ is a multiset of examples.

- (4), Naive Bayes, as an application of the Bayesian framework, is a generative approach; it models $p(\mathbf{x} \mid c)$ and $p(c)$, and hence also $p(\mathbf{x}, c)$, $p(\mathbf{x})$, and $p(c \mid \mathbf{x})$. The likelihoods, $p(\mathbf{x} \mid c)$ (or $p(x_j \mid c)$ under Naive Bayes), are estimated from $D$; the priors, $p(c)$, may be estimated by subjective assessments.

Remarks:

- Both approaches maximize $p(D)$ by maximizing $\prod_D p(c \mid \mathbf{x})$.

  Note that estimating $p(c \mid \mathbf{x})$ is usually significantly easier than estimating $p(\mathbf{x}, c)$.

- (4) Naive Bayes models $p(\mathbf{x} \mid c)$ as $\prod_{j=1}^{p} p(x_j \mid c)$, where $p(x_j \mid c)$ is estimated as $\hat{p}(x_j \mid c)$, $\hat{p}(x_j \mid c) = |\{(\mathbf{x}, c) \in D : \mathbf{x}|_j = x_j\}| \,/\, |\{(\,\cdot\,, c) \in D\}|$.

  Similarly, $p(c)$ can be estimated as $\hat{p}(c)$, $\hat{p}(c) = |\{(\,\cdot\,, c) \in D\}|$; but, also a dedicated (and subjective) prior probability model can be stated.

  $p(\mathbf{x})$ can be computed with the Law of Total Probability, $p(\mathbf{x}) = \sum_{c \in \{\oplus, \ominus\}} p(\mathbf{x} \mid c) \cdot p(c)$. Note, however, that $p(\mathbf{x})$ is not required to compute $c_{\text{MAP}}$ for $\mathbf{x}$.

- (4) If for Naive Bayes—aside from the likelihoods $p(x_j \mid c)$— also the class priors, $p(c)$, are computed from $D$, we follow the frequentist paradigm, similar to the MLE principle. Only if the values for $p(c)$ (= the prior probability model) rely on subjective assessments, the application of Naive Bayes can be considered as subjectivist.

- Whether to apply logistic regression (MLE principle) or Naive Bayes is not a free choice; it depends on

  - knowledge about the distribution of the condition events (= the hypotheses, here: $c$),
  - the distribution of feature values in the data set $D$,
  - the measurement scale of the features $x_j$.

- Synonymous: covariate, predictor, independent [variable]

Remarks: (continued)

- Observe the subtle distinction between "Bayes rule" and "Bayesian framework". With the former we refer to the identity that connects the posterior probability, $P(A \mid B)$, and the likelihood, $P(B \mid A)$ (the "reversal of condition and consequence").

  With the latter we refer to the optimization method (= comparison of possible events) where the event with the maximum a posteriori probability is determined (= MAP hypothesis). The event can be a class (as in (4)) or a distribution parameter (as in (6)).

- Note that a class-conditional event "**X**$=\mathbf{x} \mid$ *C*$=c$" does not necessarily model a cause-effect relation: the event "*C*$=c$" may cause—but does not need to cause—the event "**X**$=\mathbf{x}$".

  Examples:

  - A disease $c$ will cause the symptoms $\mathbf{x}$ (but not vice versa).
  - Weather conditions $\mathbf{x}$ will cause the decision "*EnjoySurfing*$=$yes" (but not vice versa).

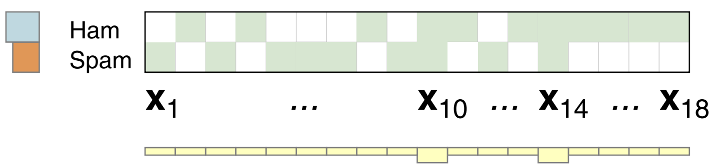
  Similarly, also if $\mathbf{x}$ is the independent variable of a function $y(\mathbf{x})$ that maps features to classes $c$, the cause-effect direction is not necessarily $\mathbf{x} \rightarrow c$, but can also be the other way around: Consider $y(\mathbf{x}) = c$ with "disease $c$" $\rightarrow$ "symptoms $\mathbf{x}$".

# Frequentist versus Subjectivist

## Logistic Regression versus Naive Bayes: Example

A multiset of examples $D$:

| | URLs | Spelling errors | Spam |
|---|---|---|---|
| 1 | 5 | 3 | yes |
| 2 | 4 | 1 | no |
| 3 | 4 | 3 | yes |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 10 | 1 | 0 | no |
| 11 | 1 | 0 | yes |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 15 | 1 | 4 | no |
| 16 | 1 | 4 | yes |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 20 | 0 | 4 | no |




Learning task:

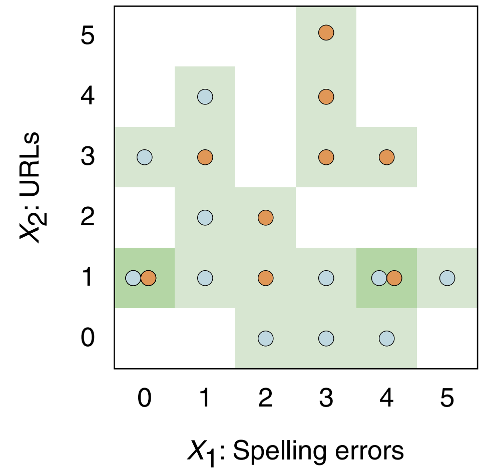
- Fit $D$ to compute a classifier for feature vectors $\mathbf{x}$, $\mathbf{x} \notin D$.

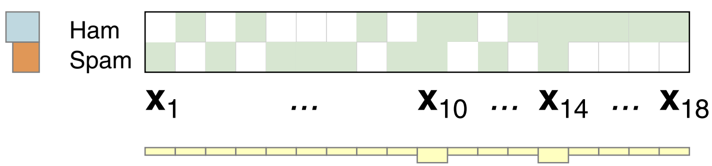
# Frequentist versus Subjectivist

## Logistic Regression versus Naive Bayes: Example

A multiset of examples $D$:

| | URLs | Spelling errors | Spam |
|---|---|---|---|
| 1 | 5 | 3 | yes |
| 2 | 4 | 1 | no |
| 3 | 4 | 3 | yes |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 10 | 1 | 0 | no |
| 11 | 1 | 0 | yes |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 15 | 1 | 4 | no |
| 16 | 1 | 4 | yes |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 20 | 0 | 4 | no |




Learning task:

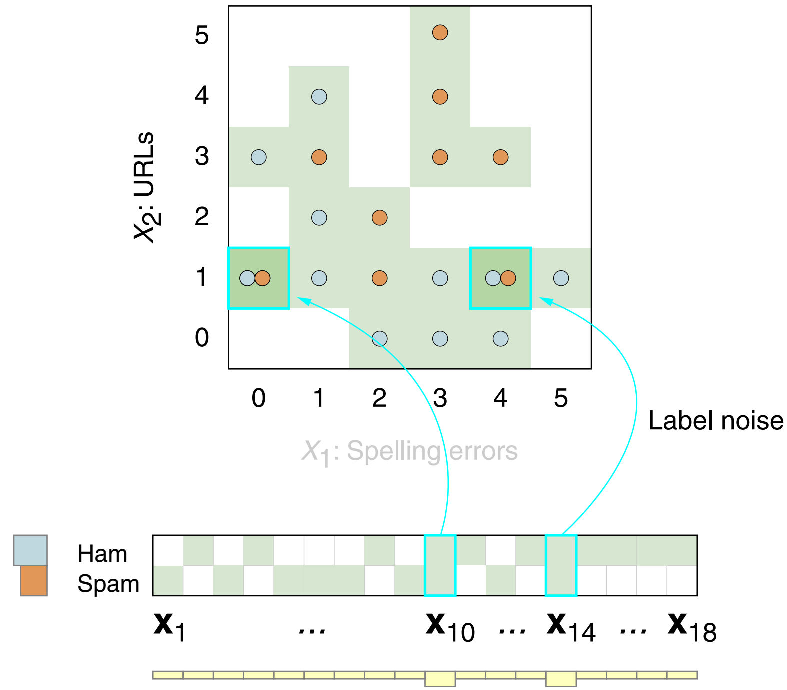
- Fit $D$ to compute a classifier for feature vectors $\mathbf{x}$, $\mathbf{x} \notin D$.

# Frequentist versus Subjectivist

## Logistic Regression versus Naive Bayes: Example

A multiset of examples $D$:

| | URLs | Spelling errors | Spam |
|---|---|---|---|
| 1 | 5 | 3 | yes |
| 2 | 4 | 1 | no |
| 3 | 4 | 3 | yes |
| ⋮ | ⋮ | ⋮ | ⋮ |
| **10** | **1** | **0** | **no** |
| **11** | **1** | **0** | **yes** |
| ⋮ | ⋮ | ⋮ | ⋮ |
| **15** | **1** | **4** | **no** |
| **16** | **1** | **4** | **yes** |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 20 | 0 | 4 | no |



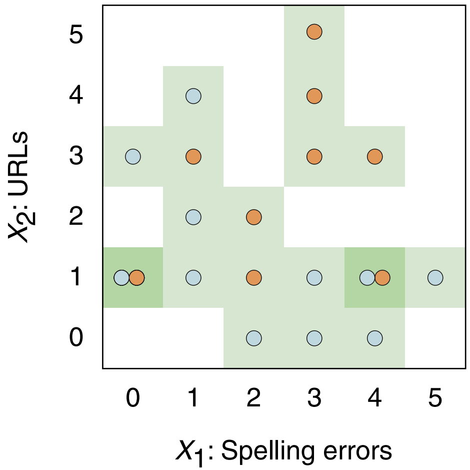

Learning task:

- Fit $D$ to compute a classifier for feature vectors $\mathbf{x}$, $\mathbf{x} \notin D$.

# Frequentist versus Subjectivist

## Logistic Regression versus Naive Bayes: Conditional Class Probabilities

Logistic regression:

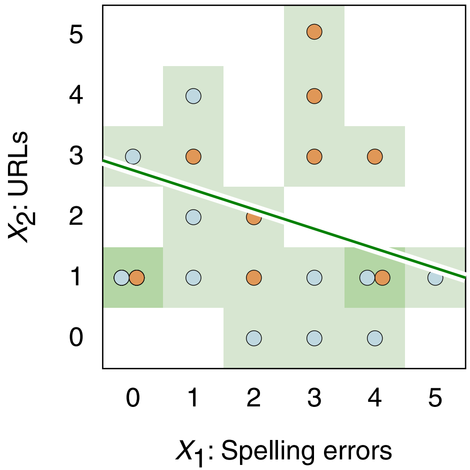



- Distribution of $D$.

# Frequentist versus Subjectivist

## Logistic Regression versus Naive Bayes: Conditional Class Probabilities

Logistic regression:

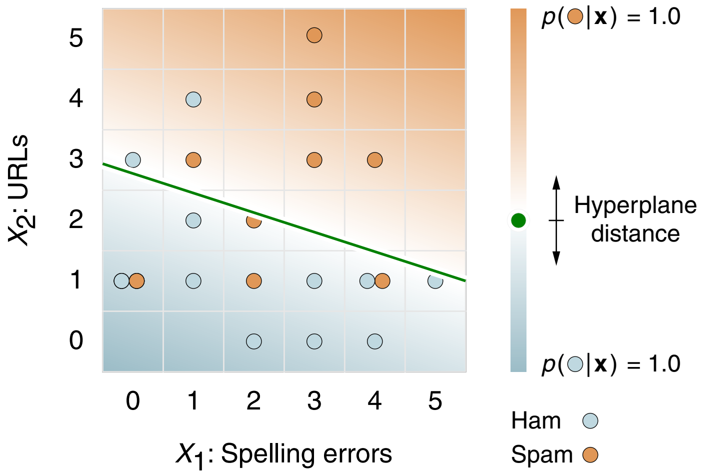



- Hyperplane $\mathbf{w}_{\text{ML}}^T \mathbf{x} = 0$. $\mathbf{w}_{\text{ML}}$ is the ML estimate for $\mathbf{w}$ given $D$.

# Frequentist versus Subjectivist

## Logistic Regression versus Naive Bayes: Conditional Class Probabilities

Logistic regression:



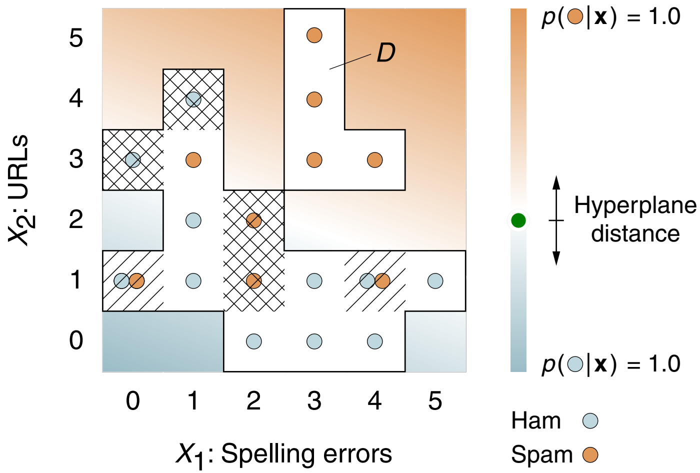

- Conditional class probabilities computed with $y(\mathbf{x}) = \sigma(\mathbf{w}_{\text{ML}}^T\mathbf{x})$.

# Frequentist versus Subjectivist

## Logistic Regression versus Naive Bayes: Conditional Class Probabilities

Logistic regression:



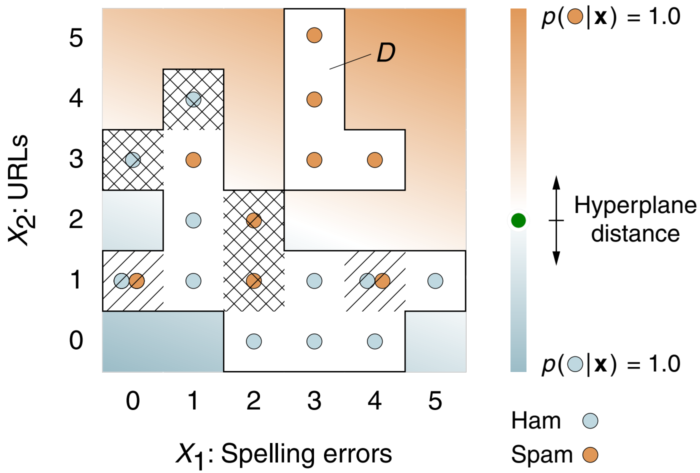

- Training error.

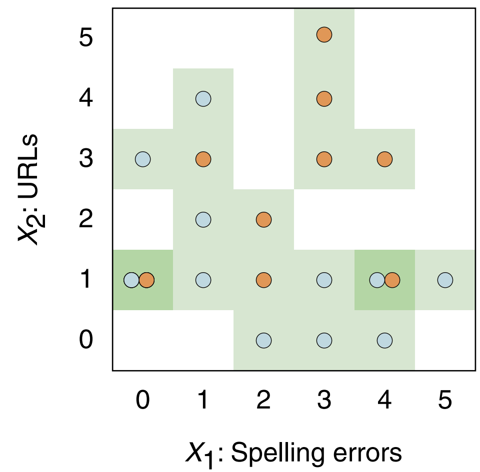
# Frequentist versus Subjectivist

## Logistic Regression versus Naive Bayes: Conditional Class Probabilities

Logistic regression:




Naive Bayes:

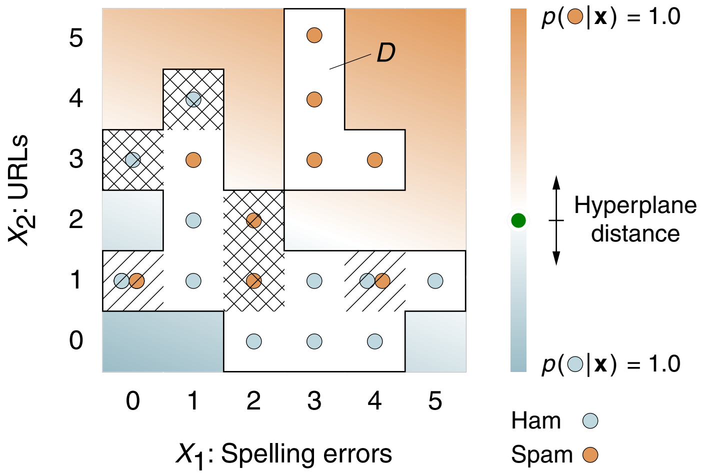



- Training error.
- Distribution of $D$.

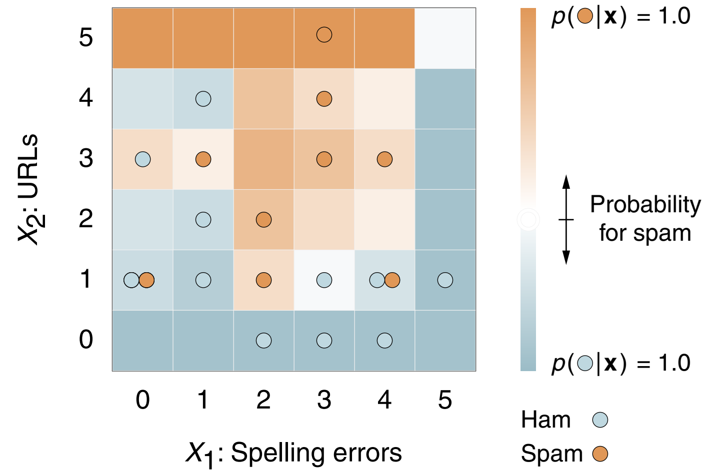
# Frequentist versus Subjectivist

## Logistic Regression versus Naive Bayes: Conditional Class Probabilities

Logistic regression:




Naive Bayes:



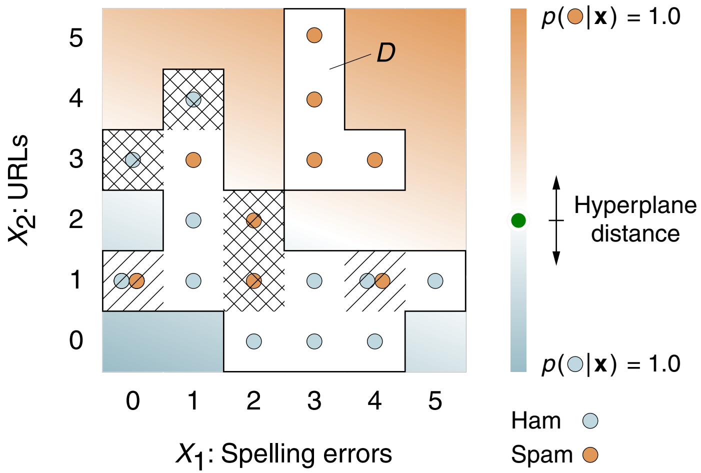

- Training error.

- Conditional class probabilities computed for the respective MAP class, using $p(c)$ estimates from $D$.

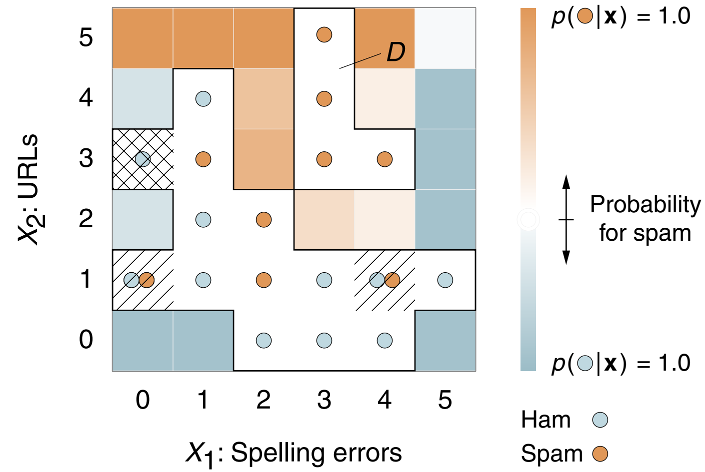
# Frequentist versus Subjectivist

## Logistic Regression versus Naive Bayes: Conditional Class Probabilities

Logistic regression:




Naive Bayes:



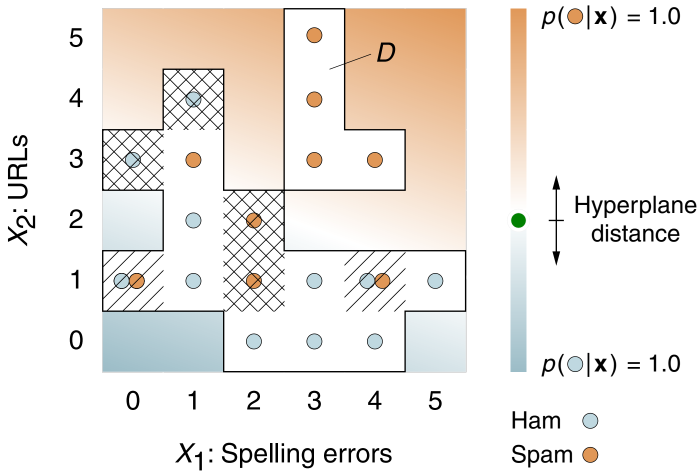

- Training error.

- Training error.

# Frequentist versus Subjectivist

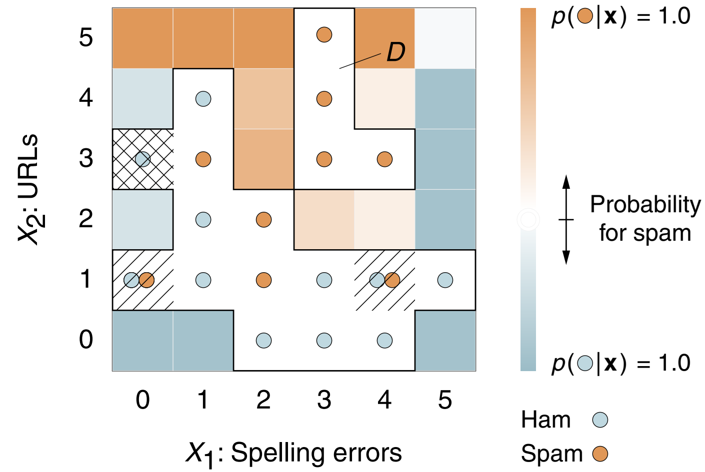
## Logistic Regression versus Naive Bayes: Conditional Class Probabilities

### Logistic regression:




- Computation of a hyperplane.
- Approach: minimization of accumulated "misclassification distances" for examples in $D$.
- Discriminative and probabilistic.

### Naive Bayes:




- Computation of a probability distribution.
- Basis: class-conditional feature and class frequencies in $D$.
- Generative (implies probabilistic).

Remarks:

- Both approaches, logistic regression and Naive Bayes, estimate the conditional class probability function, $p(\text{Spam} \mid \mathbf{x})$ or $p(\text{Ham} \mid \mathbf{x}) = 1 - p(\text{Spam} \mid \mathbf{x})$. However, the two estimation approaches follow very different concepts.

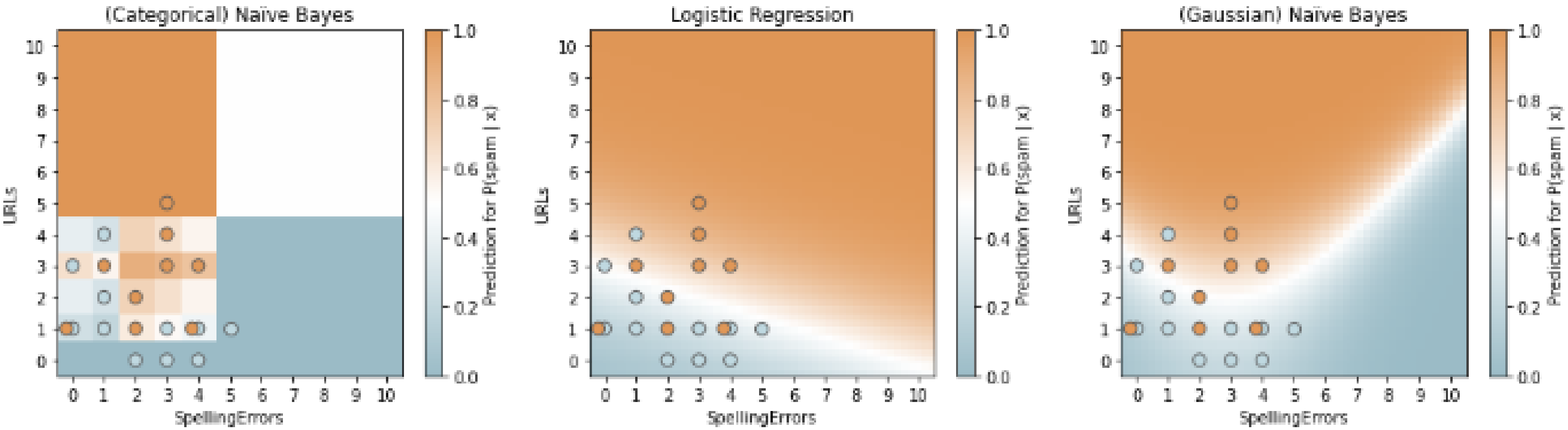
- Generalization characteristic:
  - The conditional class probability function as computed via logistic regression decides not only the feature space $\{0, 1, 2, 3, 4, 5\}^2$ but the entire $\mathbf{R}^2$ (whether this makes sense is another question).
  - The conditional class probability function as computed via Naive Bayes provides class probability estimates for $\mathbf{x} \in \{0, 1, 2, 3, 4, 5\}^2$. The probabilities are estimated from the class-conditional feature frequencies (likelihood estimates) and class frequencies, $\hat{p}(x_1 \mid c)$, $\hat{p}(x_2 \mid c)$, and $\hat{p}(c)$, as found in $D$. Note that a vector $\mathbf{x} = (x_1, x_2)^T$ gets the probability of zero for class $c$, if $x_1$ or $x_2$ does not occur in some feature vector with class label $c$ in $D$.

- Handling of class imbalance and covariate distribution:
  - Logistic regression considers the $p(c)$ and the $p(\mathbf{x})$ implicitly via their multiplicity in $D$. I.e., the learned parameter vector $\mathbf{w}_{\text{ML}}$ has the class imbalance as well as the covariate distribution "compiled in".
  - Naive Bayes, again, estimates the $p(c)$ and the $p(\mathbf{x})$ from the frequencies in $D$. More specifically, $p(\mathbf{x})$ can be estimated from $\hat{p}(x_1 \mid c)$, $\hat{p}(x_2 \mid c)$, and $\hat{p}(c)$ with the Law of Total Probability. Note that the computation of $p(\mathbf{x})$ is not necessary for a ranking (= classification without class membership probability).

# Frequentist versus Subjectivist

## Naive Bayes: Smoothing and Continuous Likelihoods




$\rightsquigarrow$ *BOARD*

# Frequentist versus Subjectivist

## Naive Bayes: Prior Probability Models

Comparison of the conditional class probability function, $p(c \mid \mathbf{x})$, under Naive Bayes for three different prior probability models (= assessments of class priors), $p(c)$.


$p(c)$ estimates from $D$

$P_a(C{=}\text{Spam}) = \hat{p}(\text{Spam}) = 0.45$

$P_a(C{=}\text{Ham}) = \hat{p}(\text{Ham}) = 0.55$

Subjective assessments for $p(c)$

$P_b(C{=}\text{Spam}) = 0.6$

$P_b(C{=}\text{Ham}) = 0.4$

$P_c(C{=}\text{Spam}) = 0.8$

$P_c(C{=}\text{Ham}) = 0.2$

# Frequentist versus Subjectivist

## Classification: Bayes Optimum versus MAP versus Ensemble

$\rightsquigarrow$ *BOARD*

# Frequentist versus Subjectivist

## Advanced Bayesian Decision Making

Recall the Bayes rule,

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)},$$

with $A$ and $B$ in the role of a "hypothesis event", $\boldsymbol{H}{=}h$, and a "data event", $\mathbf{D}{=}D$,

$$P(\boldsymbol{H}{=}h \mid \mathbf{D}{=}D) \;=\; \frac{P(\mathbf{D}{=}D \mid \boldsymbol{H}{=}h) \cdot P(\boldsymbol{H}{=}h)}{P(\mathbf{D}{=}D)}$$

# Frequentist versus Subjectivist

## Advanced Bayesian Decision Making

Recall the Bayes rule,

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)},$$

with $A$ and $B$ in the role of a "hypothesis event", $\textsf{\textit{H}}{=}h$, and a "data event", $\mathbf{D}{=}D$,

$$P(\textsf{\textit{H}}{=}h \mid \mathbf{D}{=}D) \;=\; \frac{P(\mathbf{D}{=}D \mid \textsf{\textit{H}}{=}h) \cdot P(\textsf{\textit{H}}{=}h)}{P(\mathbf{D}{=}D)}$$

rewritten using probability mass functions, pmf, (in case of discrete events) :

$$p(h \mid D) = \frac{p(D \mid h) \cdot p(h)}{p(D)}$$

- Likelihood: How well does $h$ explain (= entail, induce, evoke) the data $D$?
- Prior: How probable is the hypothesis $h$ a priori (= in principle)?
- Normalization: How probable is the observation of the data $D$?
- Posterior: How probable is the hypothesis $h$ when observing the data $D$?

# Frequentist versus Subjectivist

## Advanced Bayesian Decision Making

Recall the Bayes rule,

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)},$$

with $A$ and $B$ in the role of a "hypothesis event", $\mathsf{H}{=}h$, and a "data event", $\mathbf{D}{=}D$,

$$P(\mathsf{H}{=}h \mid \mathbf{D}{=}D) \;=\; \frac{P(\mathbf{D}{=}D \mid \mathsf{H}{=}h) \cdot P(\mathsf{H}{=}h)}{P(\mathbf{D}{=}D)}$$

rewritten using probability mass functions, pmf, (in case of discrete events) :

$$p(h \mid D) = \frac{p(D \mid h) \cdot p(h)}{p(D)}$$

- Likelihood: How well does $h$ explain (= entail, induce, evoke) the data $D$?
- Prior: How probable is the hypothesis $h$ a priori (= in principle)?
- Normalization: How probable is the observation of the data $D$?
- Posterior: How probable is the hypothesis $h$ when observing the data $D$?

# Frequentist versus Subjectivist

## Advanced Bayesian Decision Making

Recall the Bayes rule,

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)},$$

with $A$ and $B$ in the role of a "hypothesis event", $\mathbf{H}{=}h$, and a "data event", $\mathbf{D}{=}D$,

$$P(\mathbf{H}{=}h \mid \mathbf{D}{=}D) \;=\; \frac{P(\mathbf{D}{=}D \mid \mathbf{H}{=}h) \cdot P(\mathbf{H}{=}h)}{P(\mathbf{D}{=}D)}$$

rewritten using probability mass functions, pmf, (in case of discrete events) :

$$p(h \mid D) = \frac{p(D \mid h) \cdot p(h)}{p(D)}$$

- Likelihood: How well does $h$ explain (= entail, induce, evoke) the data $D$?
- Prior: How probable is the hypothesis $h$ a priori (= in principle)?
- Normalization: How probable is the observation of the data $D$?
- Posterior: How probable is the hypothesis $h$ when observing the data $D$?

# Frequentist versus Subjectivist

## Advanced Bayesian Decision Making

Recall the Bayes rule,

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)},$$

with $A$ and $B$ in the role of a "hypothesis event", $\mathbf{H}{=}h$, and a "data event", $\mathbf{D}{=}D$,

$$P(\mathbf{H}{=}h \mid \mathbf{D}{=}D) \;=\; \frac{P(\mathbf{D}{=}D \mid \mathbf{H}{=}h) \cdot P(\mathbf{H}{=}h)}{P(\mathbf{D}{=}D)}$$

rewritten using probability mass functions, pmf, (in case of discrete events) :

$$p(h \mid D) = \frac{p(D \mid h) \cdot p(h)}{p(D)}$$

- Likelihood: How well does $h$ explain (= entail, induce, evoke) the data $D$?
- Prior: How probable is the hypothesis $h$ a priori (= in principle)?
- Normalization: How probable is the observation of the data $D$?
- Posterior: How probable is the hypothesis $h$ when observing the data $D$?

Remarks:

- When using the Bayesian framework for a predictor-response setting, then $p(D)$, $p(D) := P(\mathbf{D}{=}D)$, is the probability of the data $D = \mathbf{x}$. I.e., **D** is a random vector whose domain is the feature space $\mathbf{X}$.

- When using the Bayesian framework for an outcome-only setting, then $p(D)$, $p(D) := P(\mathbf{D}{=}D)$, is the probability of the data $D = \{y_1, \ldots, y_n\}$ or $D = \{c_1, \ldots, c_n\}$. I.e., **D** is a random vector whose domain is $\mathbf{R}^n$ or $C^n$, where $C$ is the set of possible classes or class labels.

- $p(h) := P(\mathit{H}{=}h)$ (also $p(\mathbf{w})$, $p(\theta)$, or similar) is the probability of choosing a certain $h$, a parameter vector $\mathbf{w}$, or some model function as hypothesis. I.e., *H* is a random variable whose domain is the set $H$ of possible hypotheses.

- Recap. Recall that $p()$ is defined via $P()$ and that the two notations can be used interchangeably, arguing about realizations of random variables and events respectively.