

Chapter ML:VII (continued)

VII. Bayesian Learning

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

Exploitation of Data

Data Events

Data from a “predictor-response” setting:

$$D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\} \quad (\text{regression})$$

$$D = \{(\mathbf{x}_1, c_1), \dots, (\mathbf{x}_n, c_n)\} \quad (\text{classification})$$

- D is the result of n i.i.d. trials. I.e., n objects are sampled independently and from the same probability distribution. All objects are characterized by a “response” variable that is either quantitative (a number y) or categorical (a class label c), and by p “predictors” (a feature vector \mathbf{x}).
- $p(\mathbf{x}_i, c_i), p(\mathbf{x}_i, y_i) := P(\mathbf{X}_i=\mathbf{x}_i, C_i=c_i)$, is the probability of the joint event $\{\mathbf{X}_i=\mathbf{x}_i, C_i=c_i\}$, i.e., (1) to get the vector \mathbf{x}_i , and, (2) that the respective object belongs to class c_i . The $p(\mathbf{x}_i, y_i)$ are defined analogously.
- The Y_i, C_i , and \mathbf{X}_i are i.i.d. (multivariate) random variables. Typically, the Y_i are of continuous type, the C_i of discrete type, and the variables of the random vector $\mathbf{X}_i, \mathbf{X}_i := (X_{1,i}, \dots, X_{p,i})^T$, of continuous type.

Exploitation of Data

Data Events

Data from a “predictor-response” setting:

$$D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\} \quad (\text{regression})$$

$$D = \{(\mathbf{x}_1, c_1), \dots, (\mathbf{x}_n, c_n)\} \quad (\text{classification})$$

- D is the result of n i.i.d. trials. I.e., n objects are sampled independently and from the same probability distribution. All objects are characterized by a “response” variable that is either quantitative (a number y) or categorical (a class label c), and by p “predictors” (a feature vector \mathbf{x}).
- $p(\mathbf{x}_i, c_i), p(\mathbf{x}_i, y_i) := P(\mathbf{X}_i=\mathbf{x}_i, C_i=c_i)$, is the probability of the joint event $\{\mathbf{X}_i=\mathbf{x}_i, C_i=c_i\}$, i.e., (1) to get the vector \mathbf{x}_i , and, (2) that the respective object belongs to class c_i . The $p(\mathbf{x}_i, y_i)$ are defined analogously.
- The Y_i , C_i , and \mathbf{X}_i are i.i.d. (multivariate) random variables. Typically, the Y_i are of continuous type, the C_i of discrete type, and the variables of the random vector \mathbf{X}_i , $\mathbf{X}_i := (X_{1,i}, \dots, X_{p,i})^T$, of continuous type.

Exploitation of Data

Data Events (continued)

Data from an “outcome-only” setting:

$$D = \{y_1, \dots, y_n\} \quad (\text{quantitative})$$

$$D = \{c_1, \dots, c_n\} \quad (\text{categorical})$$

- D is the result of n i.i.d. trials. I.e., n outcomes are sampled independently and from the same probability distribution. All outcomes are characterized by either a number y or a class label c .
- $p(y_i), p(y_i) := P(Y_i=y_i)$, is the probability of the event $Y_i=y_i$.
 $p(c_i), p(c_i) := P(C_i=c_i)$, is the probability of the event $C_i=c_i$.
- The Y_i , and C_i are i.i.d. random variables. Typically, the Y_i are of continuous type and the C_i of discrete type.

Exploitation of Data

Data Events (continued)

Data from an “outcome-only” setting:

$$D = \{y_1, \dots, y_n\} \quad (\text{quantitative})$$

$$D = \{c_1, \dots, c_n\} \quad (\text{categorical})$$

- D is the result of n i.i.d. trials. I.e., n outcomes are sampled independently and from the same probability distribution. All outcomes are characterized by either a number y or a class label c .
- $p(y_i), p(y_i) := P(Y_i=y_i)$, is the probability of the event $Y_i=y_i$.
 $p(c_i), p(c_i) := P(C_i=c_i)$, is the probability of the event $C_i=c_i$.
- The Y_i , and C_i are i.i.d. random variables. Typically, the Y_i are of continuous type and the C_i of discrete type.

Remarks:

- The following remarks on the predictor-response setting are detailed for a categorical response variable c ; they apply to a quantitative response variable y as well.
- By [experiment design](#), the n joint events, $\{\mathbf{X}_1=\mathbf{x}_1, C_1=c_1\}, \dots, \{\mathbf{X}_n=\mathbf{x}_n, C_n=c_n\}$, generating the data D are mutually independent:

$$\begin{aligned} p(D) = p(\{(\mathbf{x}_1, c_1), \dots, (\mathbf{x}_n, c_n)\}) &= \prod_{i=1, \dots, n} p(\mathbf{x}_i, c_i) \\ &\stackrel{(1)}{=} \prod_{i=1, \dots, n} (p(\mathbf{x}_i) \cdot p(c_i \mid \mathbf{x}_i)) \\ &\stackrel{(2)}{=} \prod_{i=1, \dots, n} p(\mathbf{x}_i) \cdot \prod_{i=1, \dots, n} p(c_i \mid \mathbf{x}_i) \end{aligned}$$

- (1) Usually *not* independent are any two events $\mathbf{X}_i=\mathbf{x}_i$ and $C_i=c_i$, $i = 1, \dots, n$:

$$p(\mathbf{x}_i, c_i) \neq p(\mathbf{x}_i) \cdot p(c_i)$$

- (2) See the maximum likelihood derivation of the [logistic loss \$L_\sigma\(\mathbf{w}\)\$](#) .

- By [experiment design](#), the probabilities, $p(\mathbf{x}_i)$, $i = 1, \dots, n$, are independent, i.e., the probability of the joint event $\{\mathbf{X}_1=\mathbf{x}_1, \dots, \mathbf{X}_n=\mathbf{x}_n\}$ is equal to the product of the singleton events: $p(\mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{i=1, \dots, n} p(\mathbf{x}_i)$.

A consistent and unbiased estimate for $p(\mathbf{x})$ is $\hat{p}(\mathbf{x}) = |\{(\mathbf{x}, \cdot) \in D\}| \cdot \frac{1}{|D|}$.

- By [experiment design](#), the conditional probabilities, $p(c_i \mid \mathbf{x}_i)$, $i = 1, \dots, n$, are *invariant under covariate shift*, i.e., invariant under a change of $p(\mathbf{x}_i)$. That is, the classification procedure, “determination of c_i given some \mathbf{x}_i ”, always runs the same way, regardless of how often \mathbf{x}_i is encountered.

Remarks: (continued)

- The invariance of $p(c_i \mid \mathbf{x}_i)$ under a covariate shift can also be understood as the fact that any two events $\mathbf{X}_i = \mathbf{x}_i$ and $(C_i = c_i \mid \mathbf{X}_i = \mathbf{x}_i)$, $i = 1, \dots, n$ are independent:

$$“p(\mathbf{x}, (c \mid \mathbf{x}))” = p(\mathbf{x}) \cdot p(c \mid \mathbf{x}) = p(\mathbf{x}, c)$$

However, this interpretation is problematic since standard probability theory does not allow a conditional event being combined with other events. See section [Probability Basics](#) of this part, [conditional event algebra](#), and [Lewis’s triviality result](#) for details.

- Within an outcome-only setting such as “flipping a coin”, the object features (coin diameter, coin age, etc.) are not used as predictors. I.e., one does not model the relationship between a response variable and predictors \mathbf{x} but models (the probability of) a sequence of outcomes $D = \{y_1, \dots, y_n\}$ or $D = \{c_1, \dots, c_n\}$.
- The type of setting, be it predictor-response or outcome-only, is independent of data exploitation aspects such as
 - discriminative versus generative,
 - non-probabilistic versus probabilistic,
 - maximum likelihood versus Bayes, or
 - frequentist versus subjectivist.

Exploitation of Data

Typical Learning Settings

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

$$(1) \quad \text{RSS}(\mathbf{w}) : \quad \sum_{(\mathbf{x}, y) \in D} (y - \mathbf{w}^T \mathbf{x})^2$$

RSS for D under a linear model, parameterized by \mathbf{w} .
Least squares estimate: $\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^{p+1}} \text{RSS}(\mathbf{w})$

$$(2) \quad p(D; \mathbf{w}) : \quad \prod_{(\mathbf{x}, c) \in D} p(c \mid \mathbf{x}; \mathbf{w})$$

Probability of D under a logistic model, parameterized by \mathbf{w} . Maximum likelihood estimate:
 $\mathbf{w}_{\text{ML}} = \operatorname{argmax}_{\mathbf{w} \in \mathbb{R}^{p+1}} p(D; \mathbf{w})$

$$(3) \quad L(\mathbf{w}) : \quad \sum_{(\mathbf{x}, c) \in D} l_{\sigma}(c, \sigma(\mathbf{w}^T \mathbf{x}))$$

Loss for D under a logistic model, parameterized by \mathbf{w} .
Minimum loss (= maximum likelihood) estimate:
 $\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^{p+1}} L(\mathbf{w})$

$$(4) \quad p(c \mid \mathbf{x}) : \quad \frac{p(\mathbf{x} \mid c) \cdot p(c)}{p(\mathbf{x})}$$

Probability of c given \mathbf{x} via Bayes's rule. Maximum a posteriori class for \mathbf{x} : $c_{\text{MAP}} = \operatorname{argmax}_{c \in \{\oplus, \ominus\}} p(c \mid \mathbf{x})$

$$D = \{y_1, \dots, y_n\}, \quad D = \{c_1, \dots, c_n\}$$

$$(5) \quad p(D; \theta) : \quad \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k}$$

Probability of D under the binomial distribution, parameterized by θ . Maximum likelihood estimate:
 $\theta_{\text{ML}} = \operatorname{argmax}_{\theta \in [0;1]} p(D; \theta)$

$$(6) \quad p(\theta \mid D) : \quad \frac{p(D \mid \theta) \cdot p(\theta)}{p(D)}$$

Probability of θ given D via Bayes's rule. Maximum a posteriori hypothesis: $\theta_{\text{MAP}} = \operatorname{argmax}_{\theta \in \{\theta_1, \theta_2\}} p(\theta \mid D)$

Exploitation of Data

Typical Learning Settings

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

$$(1) \quad \text{RSS}(\mathbf{w}) : \quad \sum_{(\mathbf{x}, y) \in D} (y - \mathbf{w}^T \mathbf{x})^2$$

RSS for D under a linear model, parameterized by \mathbf{w} .
Least squares estimate: $\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^{p+1}} \text{RSS}(\mathbf{w})$

$$(2) \quad p(D; \mathbf{w}) : \quad \prod_{(\mathbf{x}, c) \in D} p(c \mid \mathbf{x}; \mathbf{w})$$

Probability of D under a logistic model, parameterized by \mathbf{w} . Maximum likelihood estimate:
 $\mathbf{w}_{\text{ML}} = \operatorname{argmax}_{\mathbf{w} \in \mathbb{R}^{p+1}} p(D; \mathbf{w})$

$$(3) \quad L(\mathbf{w}) : \quad \sum_{(\mathbf{x}, c) \in D} l_{\sigma}(c, \sigma(\mathbf{w}^T \mathbf{x}))$$

Loss for D under a logistic model, parameterized by \mathbf{w} .
Minimum loss (= maximum likelihood) estimate:
 $\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^{p+1}} L(\mathbf{w})$

$$(4) \quad p(c \mid \mathbf{x}) : \quad \frac{p(\mathbf{x} \mid c) \cdot p(c)}{p(\mathbf{x})}$$

Probability of c given \mathbf{x} via Bayes's rule. Maximum a posteriori class for \mathbf{x} : $c_{\text{MAP}} = \operatorname{argmax}_{c \in \{\oplus, \ominus\}} p(c \mid \mathbf{x})$

$$D = \{y_1, \dots, y_n\}, \quad D = \{c_1, \dots, c_n\}$$

$$(5) \quad p(D; \theta) : \quad \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k}$$

Probability of D under the binomial distribution, parameterized by θ . Maximum likelihood estimate:
 $\theta_{\text{ML}} = \operatorname{argmax}_{\theta \in [0;1]} p(D; \theta)$

$$(6) \quad p(\theta \mid D) : \quad \frac{p(D \mid \theta) \cdot p(\theta)}{p(D)}$$

Probability of θ given D via Bayes's rule. Maximum a posteriori hypothesis: $\theta_{\text{MAP}} = \operatorname{argmax}_{\theta \in \{\theta_1, \theta_2\}} p(\theta \mid D)$

Exploitation of Data

Typical Learning Settings

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

$$(1) \quad \text{RSS}(\mathbf{w}) : \quad \sum_{(\mathbf{x}, y) \in D} (y - \mathbf{w}^T \mathbf{x})^2$$

RSS for D under a linear model, parameterized by \mathbf{w} .
Least squares estimate: $\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^{p+1}} \text{RSS}(\mathbf{w})$

$$(2) \quad p(D; \mathbf{w}) : \quad \prod_{(\mathbf{x}, c) \in D} p(c \mid \mathbf{x}; \mathbf{w})$$

Probability of D under a logistic model, parameterized by \mathbf{w} . Maximum likelihood estimate:
 $\mathbf{w}_{\text{ML}} = \operatorname{argmax}_{\mathbf{w} \in \mathbb{R}^{p+1}} p(D; \mathbf{w})$

$$(3) \quad L(\mathbf{w}) : \quad \sum_{(\mathbf{x}, c) \in D} l_{\sigma}(c, \sigma(\mathbf{w}^T \mathbf{x}))$$

Loss for D under a logistic model, parameterized by \mathbf{w} .
Minimum loss (= maximum likelihood) estimate:
 $\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^{p+1}} L(\mathbf{w})$

$$(4) \quad p(c \mid \mathbf{x}) : \quad \frac{p(\mathbf{x} \mid c) \cdot p(c)}{p(\mathbf{x})}$$

Probability of c given \mathbf{x} via Bayes's rule. Maximum a posteriori class for \mathbf{x} : $c_{\text{MAP}} = \operatorname{argmax}_{c \in \{\oplus, \ominus\}} p(c \mid \mathbf{x})$

$$D = \{y_1, \dots, y_n\}, \quad D = \{c_1, \dots, c_n\}$$

$$(5) \quad p(D; \theta) : \quad \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k}$$

Probability of D under the binomial distribution, parameterized by θ . Maximum likelihood estimate:
 $\theta_{\text{ML}} = \operatorname{argmax}_{\theta \in [0;1]} p(D; \theta)$

$$(6) \quad p(\theta \mid D) : \quad \frac{p(D \mid \theta) \cdot p(\theta)}{p(D)}$$

Probability of θ given D via Bayes's rule. Maximum a posteriori hypothesis: $\theta_{\text{MAP}} = \operatorname{argmax}_{\theta \in \{\theta_1, \theta_2\}} p(\theta \mid D)$

Exploitation of Data

Typical Learning Settings

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

$$(1) \quad \text{RSS}(\mathbf{w}) : \quad \sum_{(\mathbf{x}, y) \in D} (y - \mathbf{w}^T \mathbf{x})^2$$

RSS for D under a linear model, parameterized by \mathbf{w} .
Least squares estimate: $\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^{p+1}} \text{RSS}(\mathbf{w})$

$$(2) \quad p(D; \mathbf{w}) : \quad \prod_{(\mathbf{x}, c) \in D} p(c \mid \mathbf{x}; \mathbf{w})$$

Probability of D under a logistic model, parameterized by \mathbf{w} . Maximum likelihood estimate:
 $\mathbf{w}_{\text{ML}} = \operatorname{argmax}_{\mathbf{w} \in \mathbb{R}^{p+1}} p(D; \mathbf{w})$

$$(3) \quad L(\mathbf{w}) : \quad \sum_{(\mathbf{x}, c) \in D} l_{\sigma}(c, \sigma(\mathbf{w}^T \mathbf{x}))$$

Loss for D under a logistic model, parameterized by \mathbf{w} .
Minimum loss (= maximum likelihood) estimate:
 $\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^{p+1}} L(\mathbf{w})$

$$(4) \quad p(c \mid \mathbf{x}) : \quad \frac{p(\mathbf{x} \mid c) \cdot p(c)}{p(\mathbf{x})}$$

Probability of c given \mathbf{x} via Bayes's rule. Maximum a posteriori class for \mathbf{x} : $c_{\text{MAP}} = \operatorname{argmax}_{c \in \{\oplus, \ominus\}} p(c \mid \mathbf{x})$

$$D = \{y_1, \dots, y_n\}, \quad D = \{c_1, \dots, c_n\}$$

$$(5) \quad p(D; \theta) : \quad \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k}$$

Probability of D under the binomial distribution, parameterized by θ . Maximum likelihood estimate:
 $\theta_{\text{ML}} = \operatorname{argmax}_{\theta \in [0;1]} p(D; \theta)$

$$(6) \quad p(\theta \mid D) : \quad \frac{p(D \mid \theta) \cdot p(\theta)}{p(D)}$$

Probability of θ given D via Bayes's rule. Maximum a posteriori hypothesis: $\theta_{\text{MAP}} = \operatorname{argmax}_{\theta \in \{\theta_1, \theta_2\}} p(\theta \mid D)$

Exploitation of Data

Typical Learning Settings

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

$$(1) \quad \text{RSS}(\mathbf{w}) : \quad \sum_{(\mathbf{x}, y) \in D} (y - \mathbf{w}^T \mathbf{x})^2$$

RSS for D under a linear model, parameterized by \mathbf{w} .
Least squares estimate: $\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^{p+1}} \text{RSS}(\mathbf{w})$

$$(2) \quad p(D; \mathbf{w}) : \quad \prod_{(\mathbf{x}, c) \in D} p(c \mid \mathbf{x}; \mathbf{w})$$

Probability of D under a logistic model, parameterized by \mathbf{w} . Maximum likelihood estimate:
 $\mathbf{w}_{\text{ML}} = \operatorname{argmax}_{\mathbf{w} \in \mathbb{R}^{p+1}} p(D; \mathbf{w})$

$$(3) \quad L(\mathbf{w}) : \quad \sum_{(\mathbf{x}, c) \in D} l_\sigma(c, \sigma(\mathbf{w}^T \mathbf{x}))$$

Loss for D under a logistic model, parameterized by \mathbf{w} .
Minimum loss (= maximum likelihood) estimate:
 $\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^{p+1}} L(\mathbf{w})$

$$(4) \quad p(c \mid \mathbf{x}) : \quad \frac{p(\mathbf{x} \mid c) \cdot p(c)}{p(\mathbf{x})}$$

Probability of c given \mathbf{x} via Bayes's rule. Maximum a posteriori class for \mathbf{x} : $c_{\text{MAP}} = \operatorname{argmax}_{c \in \{\oplus, \ominus\}} p(c \mid \mathbf{x})$

$$D = \{y_1, \dots, y_n\}, \quad D = \{c_1, \dots, c_n\}$$

$$(5) \quad p(D; \theta) : \quad \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k}$$

Probability of D under the binomial distribution, parameterized by θ . Maximum likelihood estimate:
 $\theta_{\text{ML}} = \operatorname{argmax}_{\theta \in [0;1]} p(D; \theta)$

$$(6) \quad p(\theta \mid D) : \quad \frac{p(D \mid \theta) \cdot p(\theta)}{p(D)}$$

Probability of θ given D via Bayes's rule. Maximum a posteriori hypothesis: $\theta_{\text{MAP}} = \operatorname{argmax}_{\theta \in \{\theta_1, \theta_2\}} p(\theta \mid D)$

Exploitation of Data

Typical Learning Settings

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

$$(1) \quad \text{RSS}(\mathbf{w}) : \quad \sum_{(\mathbf{x}, y) \in D} (y - \mathbf{w}^T \mathbf{x})^2$$

RSS for D under a linear model, parameterized by \mathbf{w} .
Least squares estimate: $\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^{p+1}} \text{RSS}(\mathbf{w})$

$$(2) \quad p(D; \mathbf{w}) : \quad \prod_{(\mathbf{x}, c) \in D} p(c \mid \mathbf{x}; \mathbf{w})$$

Probability of D under a logistic model, parameterized by \mathbf{w} . Maximum likelihood estimate:
 $\mathbf{w}_{\text{ML}} = \operatorname{argmax}_{\mathbf{w} \in \mathbb{R}^{p+1}} p(D; \mathbf{w})$

$$(3) \quad L(\mathbf{w}) : \quad \sum_{(\mathbf{x}, c) \in D} l_\sigma(c, \sigma(\mathbf{w}^T \mathbf{x}))$$

Loss for D under a logistic model, parameterized by \mathbf{w} .
Minimum loss (= maximum likelihood) estimate:
 $\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^{p+1}} L(\mathbf{w})$

$$(4) \quad p(c \mid \mathbf{x}) : \quad \frac{p(\mathbf{x} \mid c) \cdot p(c)}{p(\mathbf{x})}$$

Probability of c given \mathbf{x} via Bayes's rule. Maximum a posteriori class for \mathbf{x} : $c_{\text{MAP}} = \operatorname{argmax}_{c \in \{\oplus, \ominus\}} p(c \mid \mathbf{x})$

$$D = \{y_1, \dots, y_n\}, \quad D = \{c_1, \dots, c_n\}$$

$$(5) \quad p(D; \theta) : \quad \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k}$$

Probability of D under the binomial distribution, parameterized by θ . Maximum likelihood estimate:
 $\theta_{\text{ML}} = \operatorname{argmax}_{\theta \in [0;1]} p(D; \theta)$

$$(6) \quad p(\theta \mid D) : \quad \frac{p(D \mid \theta) \cdot p(\theta)}{p(D)}$$

Probability of θ given D via Bayes's rule. Maximum a posteriori hypothesis: $\theta_{\text{MAP}} = \operatorname{argmax}_{\theta \in \{\theta_1, \theta_2\}} p(\theta \mid D)$

Exploitation of Data

Typical Learning Settings

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

$$(1) \quad \text{RSS}(\mathbf{w}) : \quad \sum_{(\mathbf{x}, y) \in D} (y - \mathbf{w}^T \mathbf{x})^2$$

RSS for D under a linear model, parameterized by \mathbf{w} .
Least squares estimate: $\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^{p+1}} \text{RSS}(\mathbf{w})$

$$(2) \quad p(D; \mathbf{w}) : \quad \prod_{(\mathbf{x}, c) \in D} p(c \mid \mathbf{x}; \mathbf{w})$$

Probability of D under a logistic model, parameterized by \mathbf{w} . Maximum likelihood estimate:
 $\mathbf{w}_{\text{ML}} = \operatorname{argmax}_{\mathbf{w} \in \mathbb{R}^{p+1}} p(D; \mathbf{w})$

$$(3) \quad L(\mathbf{w}) : \quad \sum_{(\mathbf{x}, c) \in D} l_{\sigma}(c, \sigma(\mathbf{w}^T \mathbf{x}))$$

Loss for D under a logistic model, parameterized by \mathbf{w} .
Minimum loss (= maximum likelihood) estimate:
 $\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^{p+1}} L(\mathbf{w})$

$$(4) \quad p(c \mid \mathbf{x}) : \quad \frac{p(\mathbf{x} \mid c) \cdot p(c)}{p(\mathbf{x})}$$

Probability of c given \mathbf{x} via Bayes's rule. Maximum a posteriori class for \mathbf{x} : $c_{\text{MAP}} = \operatorname{argmax}_{c \in \{\oplus, \ominus\}} p(c \mid \mathbf{x})$

$$D = \{y_1, \dots, y_n\}, \quad D = \{c_1, \dots, c_n\}$$

$$(5) \quad p(D; \theta) : \quad \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k}$$

Probability of D under the binomial distribution, parameterized by θ . Maximum likelihood estimate:
 $\theta_{\text{ML}} = \operatorname{argmax}_{\theta \in [0;1]} p(D; \theta)$

$$(6) \quad p(\theta \mid D) : \quad \frac{p(D \mid \theta) \cdot p(\theta)}{p(D)}$$

Probability of θ given D via Bayes's rule. Maximum a posteriori hypothesis: $\theta_{\text{MAP}} = \operatorname{argmax}_{\theta \in \{\theta_1, \theta_2\}} p(\theta \mid D)$

Remarks (predictor-response vs. outcome-only setting) :

- (1), ..., (4) Predictor-response setting, $\mathbf{x} \rightarrow y$ or $\mathbf{x} \rightarrow c$. The relation between \mathbf{x} and y or c is captured by a model function $f(\mathbf{x})$. The data D is exploited to fit $f(\mathbf{x})$, which in turn means to determine a parameter w or parameter vector \mathbf{w} for $f(\mathbf{x})$. Modeling and predicting a quantitative response variable y is a regression task; modeling and predicting a categorical response variable c is a classification task.

An example for a categorical predictor-response setting is the classification of an email as spam ($c = \oplus$) or ham ($c = \ominus$), given a vector \mathbf{x} of linguistic features for that email.

- (5), (6) Outcome-only setting, y_1, \dots, y_n or c_1, \dots, c_n . Modeling a sole outcome variable means to fit the data D using a suited distribution function, which in turn means to determine the distribution parameter θ or distribution parameters $\boldsymbol{\theta}$. Again, one can distinguish between different measurement scales, such as quantitative (y) or categorical (c).

An example for a categorical outcome-only setting is a coin flip experiment where one has to fit the observations (number of heads and tails) under the binomial distribution, which in turn means to determine the distribution parameter θ .

- (1), ..., (6) Depending on the experiment setting, i.e., fitting of a model function vs. fitting of a distribution, either the symbol w (or \mathbf{w}), or the symbol θ (or $\boldsymbol{\theta}$) may be used to denote the parameter (or parameter vector).

Remarks (discriminative vs. generative approach) :

- (1), (2), (3) Discriminative approach to classification. Exploit the data to determine a decision boundary. Typically, “discriminative” implies “frequentist”.

The optimization (argmin, argmax) 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.

- (2), (3), (5) Maximum likelihood (ML) principle to parameter estimation.

- (2) Recall the identities from the maximum likelihood derivation of the logistic loss $L_\sigma(\mathbf{w})$:

$$p(D; \mathbf{w}) = \prod_{(\mathbf{x}, c) \in D} p(\mathbf{x}, c; \mathbf{w}), \quad \operatorname{argmax}_{\mathbf{w} \in \mathbf{R}^{p+1}} p(D; \mathbf{w}) = \operatorname{argmax}_{\mathbf{w} \in \mathbf{R}^{p+1}} \prod_{(\mathbf{x}, c) \in D} p(c \mid \mathbf{x}; \mathbf{w})$$

- (1), (2) If the data comes from an exponential family and mild conditions are satisfied, least-squares estimates and maximum-likelihood estimates are identical.

- (2), (3) Probabilistic model. The conditional class probability function (CCPF), $p(c \mid \mathbf{x})$, is estimated for all feature vectors (= at all quantiles). The model is not generative since the distribution of the independent variable, $p(\mathbf{x})$, is not modeled (but of course exploited implicitly via D).

Maximizing the probability under a logistic model is equivalent to minimizing the logistic loss L_σ . Hence, $\mathbf{w}_{\text{ML}} = \hat{\mathbf{w}}$.

Remarks (discriminative vs. generative approach) : (continued)

- (4) Generative approach to classification. Exploit the data D (here: estimate $p(\mathbf{x} \mid c)$ and $p(c)$ for all \mathbf{x} and c) to provide a model for the joint probability distribution, $p(\mathbf{x}, c)$, from which D is sampled.
- (5) Generative approach. Assuming the conditions of the binomial data model, exploit the data D (here: estimate the parameter θ) to provide a model for the binomial probability distribution, $p(c)$, from which D is sampled.
- (6) Generative or discriminative approach. $p(\theta \mid D)$ can be estimated by either providing (\rightarrow generative) or by *not* providing (\rightarrow discriminative) a model for the probability distribution from which D is sampled.

Remarks (ML principle vs. Bayes method) :

- (1), (2), (3) \mathbf{w} (as well as θ) is not the realization of a random variable—which would come along with
(5) a distribution—but an *exogenous parameter*, which is varied in order to find the maximum probability $p(D; \mathbf{w})$ (or $p(D; \theta)$ or the minimum loss $L(\mathbf{w})$).

The fact that \mathbf{w} (or θ) is an exogenous parameter and not a realization of a random variable is reflected by the notation, which uses a $\gg; \ll$ instead of a $\gg| \ll$ in $p(\cdot)$.

- (4) Application of Bayes's rule, presupposing that one can estimate the likelihoods $p(\mathbf{x} | \cdot)$ ($p(x_j | \cdot)$ in case of Naive Bayes) at higher fidelity than the conditional class probabilities, $p(\cdot | \mathbf{x})$, from the data.

Under the Naive Bayes Assumption, $p(\mathbf{x} | c)$ is modeled as $\prod_{j=1}^p p(x_j | c)$.

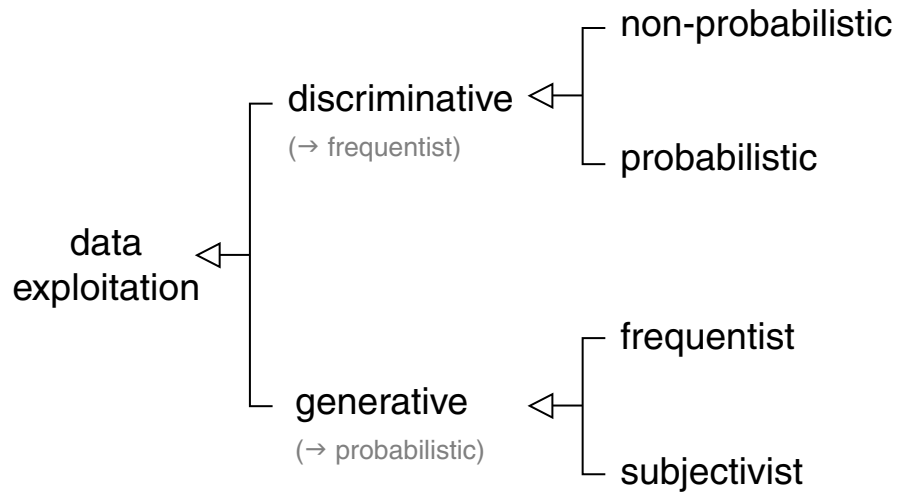
- (4), (6) Likelihoods, $p(\mathbf{x} | \cdot)$, $p(D | \cdot)$, are computed for events under alternative classes c or parameters θ . The settings differ in that an event in (4) is about a feature vector \mathbf{x} , while an event in (6) is about a sequence D . (4) may (but not need to) apply the Naive Bayes assumption to compute the likelihood $p(\mathbf{x} | c)$, which is a common approximation for a nominal feature space and if data are sparse. For (6), if the data originate from a coin flip experiment, the likelihood $p(D | \theta)$ is computed via the binomial distribution.

If the prior probabilities, $p(c)$ or $p(\theta)$, are estimated also from D , we follow the frequentist paradigm; if the priors rely on subjective assessments we follow the subjectivist paradigm.

If we assume uniform priors, i.e., the $p(c)$ or the $p(\theta)$ are equally probable, MAP estimates and ML estimates are equal since $p(c | \mathbf{x}) \propto p(\mathbf{x} | c)$ or $p(\theta | D) \propto p(D | \theta)$, where \propto means “is proportional to”.

Exploitation of Data

Learning Approaches Overview



Support vector machine

- (1) Linear regression with least square estimates from D
- (2) Logistic regression via $p(\cdot)$ with ML estimates from D
- (3) Logistic regression via $L(\cdot)$ with ML estimates from D

(4) Bayes with ML estimates from D as priors

(5) Probability model with ML estimate from D

(6) Bayes with subjective priors

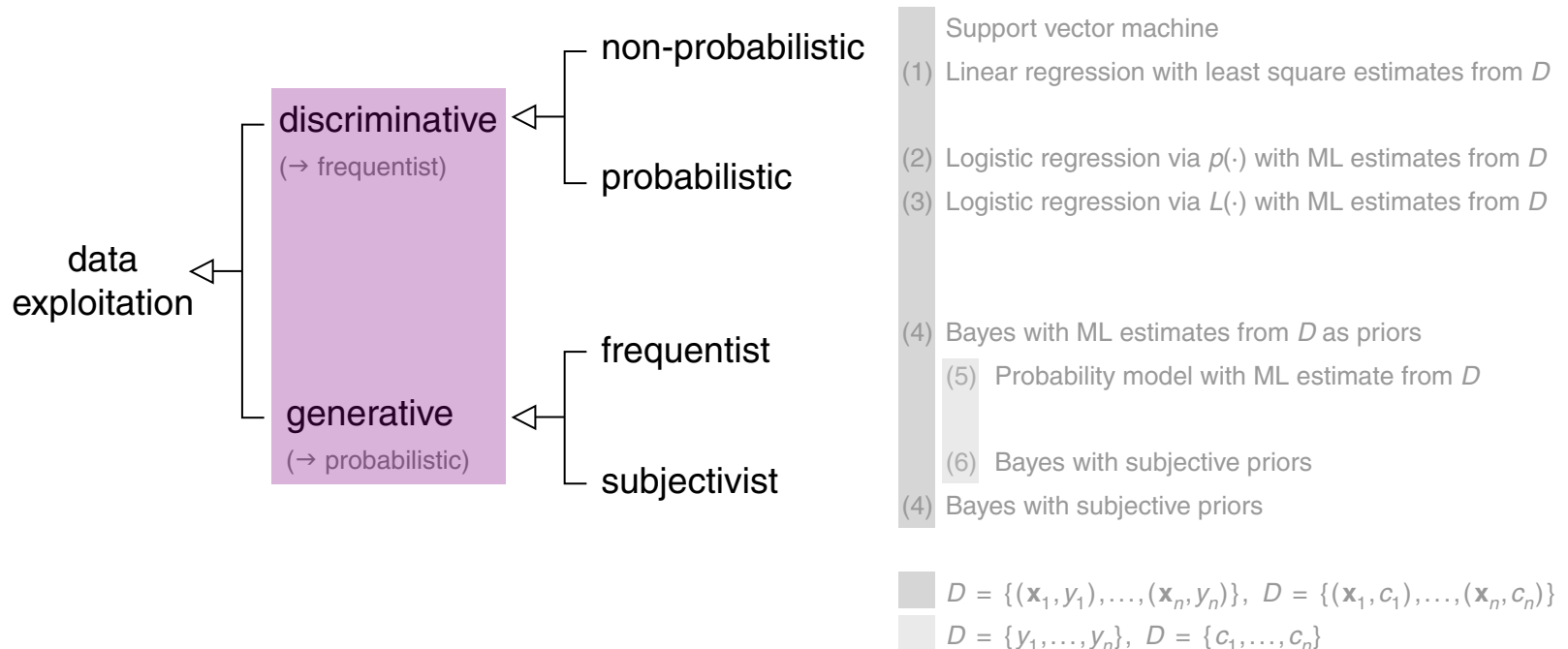
(4) Bayes with subjective priors

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

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

Exploitation of Data

Learning Approaches Overview (continued)

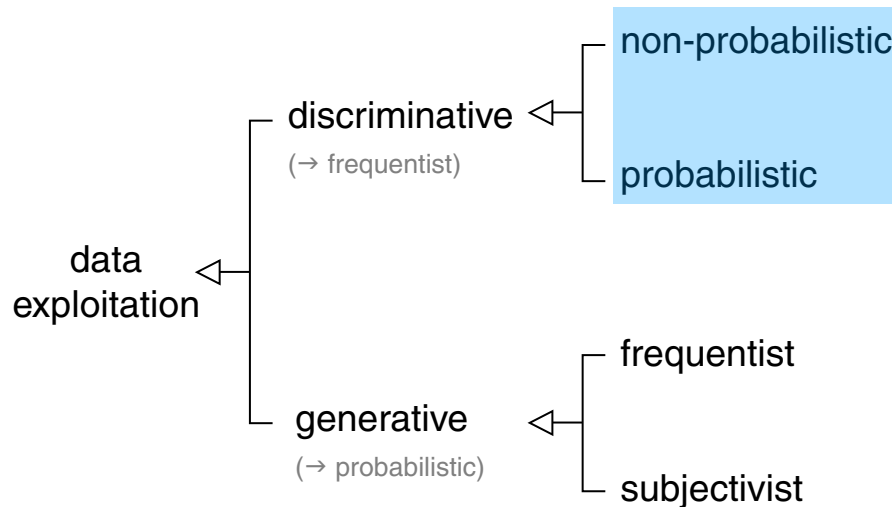


discriminative : Determine a boundary to split D . → No model for the distribution of D .

generative : Provide a model for the probability distribution from which D is sampled.

Exploitation of Data

Learning Approaches Overview (continued)



- Support vector machine
- (1) Linear regression with least square estimates from D
- (2) Logistic regression via $p(\cdot)$ with ML estimates from D
- (3) Logistic regression via $L(\cdot)$ with ML estimates from D
- (4) Bayes with ML estimates from D as priors
- (5) Probability model with ML estimate from D
- (6) Bayes with subjective priors
- (4) Bayes with subjective priors

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

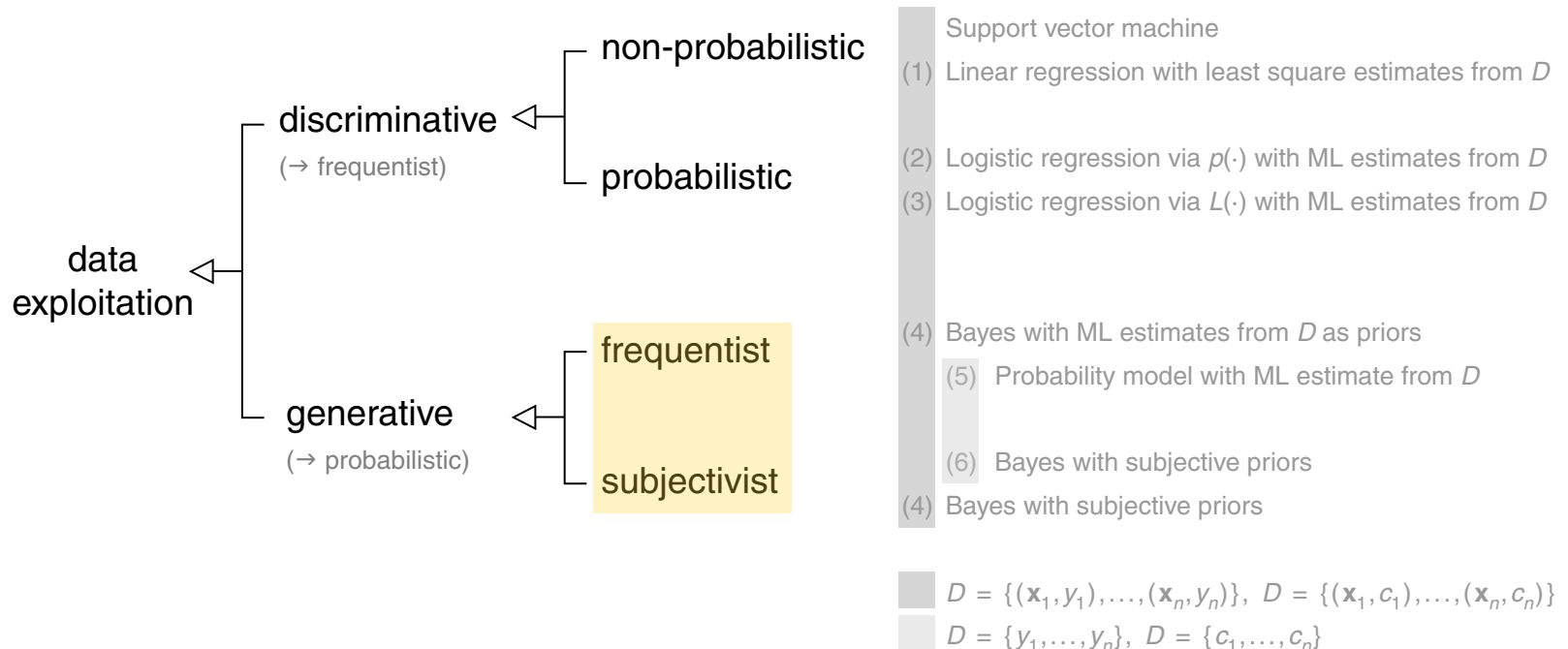
$$D = \{y_1, \dots, y_n\}, \quad D = \{c_1, \dots, c_n\}$$

non-probabilistic: Threshold some model function (typically at zero). → Classification, Labeling

probabilistic: Estimate $p(c \mid \mathbf{x})$ at all quantiles. → Class probability estimation, CCPF

Exploitation of Data

Learning Approaches Overview (continued)



frequentist: Consider a unique mechanism that generated the data D .

subjectivist: Specify beliefs for alternative mechanisms one of which generated D .

Remarks:

- ❑ We call a data exploitation approach “generative” if it provides us with a model for the probability distribution from which D is sampled. With such a model we are able to generate arbitrary samples from the population where D is sampled from.
- ❑ The overview does not show all but common combinations. In particular:
 - Typically, “discriminative” implies “frequentist”. The inverse does not apply: consider a Bayes classifier with priors estimated from the data.
 - Typically, “generative” implies “probabilistic”. The inverse does not apply: logistic regression provides a probabilistic model to classification.
- ❑ Discriminative approaches are further distinguished as “non-probabilistic” or “probabilistic”.
- ❑ Generative approaches are further distinguished as “frequentist” or “subjectivist”.

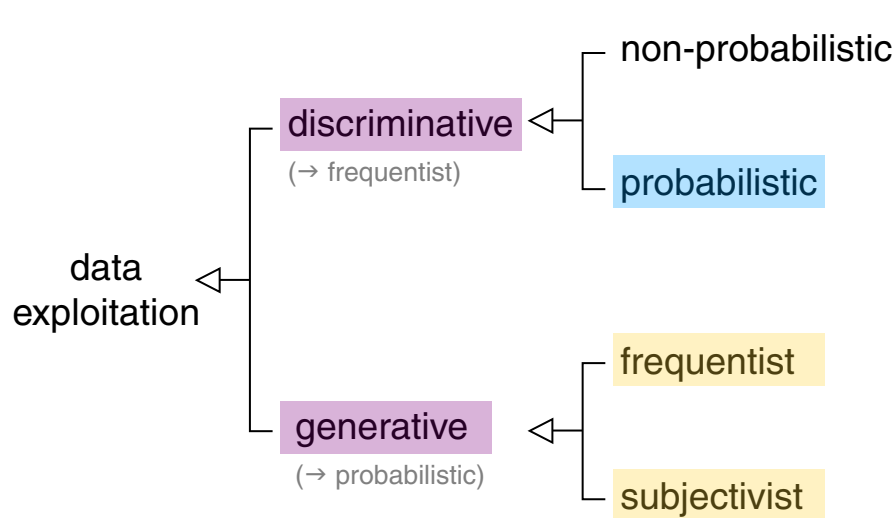
Chapter ML:VII (continued)

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 $p(\cdot)$ with ML estimates from D
- (3) Logistic regression via $L(\cdot)$ with ML estimates from D

- (4) Bayes with ML estimates from D as priors

- (5) Probability model with ML estimate from D

- (6) Bayes with subjective priors

- (4) Bayes with subjective priors

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

$$D = \{y_1, \dots, y_n\}, \quad D = \{c_1, \dots, c_n\}$$

Frequentist versus Subjectivist

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

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

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

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

- (2), the “ML 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), the “Bayes method”, determines for a given \mathbf{x} its most probable class. By choosing c_{MAP} for each \mathbf{x} , Bayes maximizes $p(D)$ by maximizing each factor of $\prod_D p(c \mid \mathbf{x})$. Note that $p(\mathbf{x})$ is constant per factor, and that 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 [\[data exploitation examples\]](#) (continued)

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

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

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

- (2), the “ML 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), the “Bayes method”, determines for a given \mathbf{x} its most probable class. By choosing c_{MAP} for each \mathbf{x} , Bayes maximizes $p(D)$ by maximizing each factor of $\prod_D p(c \mid \mathbf{x})$. Note that $p(\mathbf{x})$ is constant per factor, and that 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 [\[data exploitation examples\]](#) (continued)

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

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

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

- (2), the “ML 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), the “Bayes method”, determines for a given \mathbf{x} its most probable class. By choosing c_{MAP} for each \mathbf{x} , Bayes maximizes $p(D)$ by maximizing each factor of $\prod_D p(c \mid \mathbf{x})$. Note that $p(\mathbf{x})$ is constant per factor, and that 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 [\[data exploitation examples\]](#) (continued)

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

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

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

- (2), the “ML 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), the “Bayes method”, determines for a given \mathbf{x} its most probable class. By choosing c_{MAP} for each \mathbf{x} , Bayes maximizes $p(D)$ by maximizing each factor of $\prod_D p(c \mid \mathbf{x})$. Note that $p(\mathbf{x})$ is constant per factor, and that 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 [\[data exploitation examples\]](#) (continued)

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

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

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

- (2), the “ML 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), the “Bayes method”, determines for a given \mathbf{x} its most probable class. By choosing c_{MAP} for each \mathbf{x} , Bayes maximizes $p(D)$ by maximizing each factor of $\prod_D p(c \mid \mathbf{x})$. Note that $p(\mathbf{x})$ is constant per factor, and that 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 [\[data exploitation examples\]](#) (continued)

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

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

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

- (2), the “ML 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), the “Bayes method”, as a generative approach, 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 subjective.

Remarks:

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

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_{MAP} for \mathbf{x} .

- (4) If for the Bayes method—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 ML principle. Only if the values for $p(c)$ (= the prior probability model) rely on subjective assessments, the Bayes method can be considered as subjectivist.

- Whether to apply the ML principle or the Bayes method is not a free choice; it depends on
 - the availability of data D ,
 - the conditional strengths of the likelihoods, $p(\mathbf{x} \mid c)$, and
 - the reliability of the assessments for the prior probabilities, $p(c)$.
- Synonymous: covariate, independent, predictor (variable / distribution).

Remarks: (continued)

- Observe the subtle distinction between “Bayes rule” and “Bayes method” made here. 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 *parameter estimation principle* where the maximum a posteriori probability is determined.
- Note that a class-conditional event “ $\mathbf{X}=\mathbf{x} \mid C=c$ ” does not necessarily model a cause-effect relation: the event “ $C=c$ ” may cause—but does not need to cause—the event “ $\mathbf{X}=\mathbf{x}$ ”.

Examples:

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

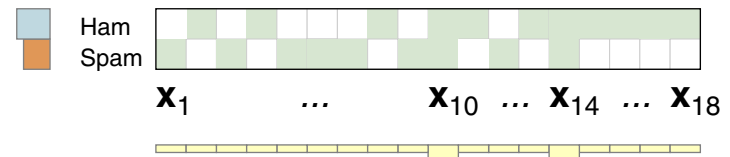
Similarly, also if \mathbf{x} is the independent variable of a function $f(\mathbf{x})$ that maps features to classes c , the cause-effect direction is not necessarily $\mathbf{x} \rightarrow c$, but can also be the other way around: Consider $f(\mathbf{x}) = c$ with “disease c ” \rightarrow “symptoms \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
\vdots	\vdots	\vdots	\vdots
10	1	0	no
11	1	0	yes
\vdots	\vdots	\vdots	\vdots
15	1	4	no
16	1	4	yes
\vdots	\vdots	\vdots	\vdots
20	0	4	no



Learning task:

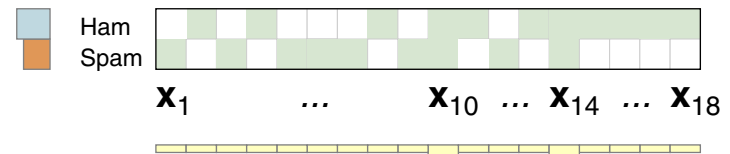
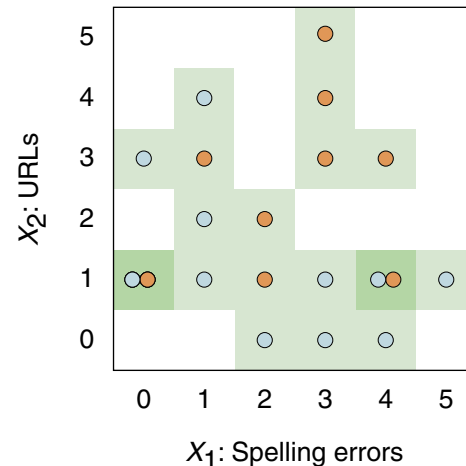
- Approximate D to compute a classifier for feature vectors x , $x \notin D$.

Frequentist versus Subjectivist

Logistic Regression versus Naive Bayes: Example (continued)

A multiset of examples D :

	URLs	Spelling errors	Spam
1	5	3	yes
2	4	1	no
3	4	3	yes
\vdots	\vdots	\vdots	\vdots
10	1	0	no
11	1	0	yes
\vdots	\vdots	\vdots	\vdots
15	1	4	no
16	1	4	yes
\vdots	\vdots	\vdots	\vdots
20	0	4	no



Learning task:

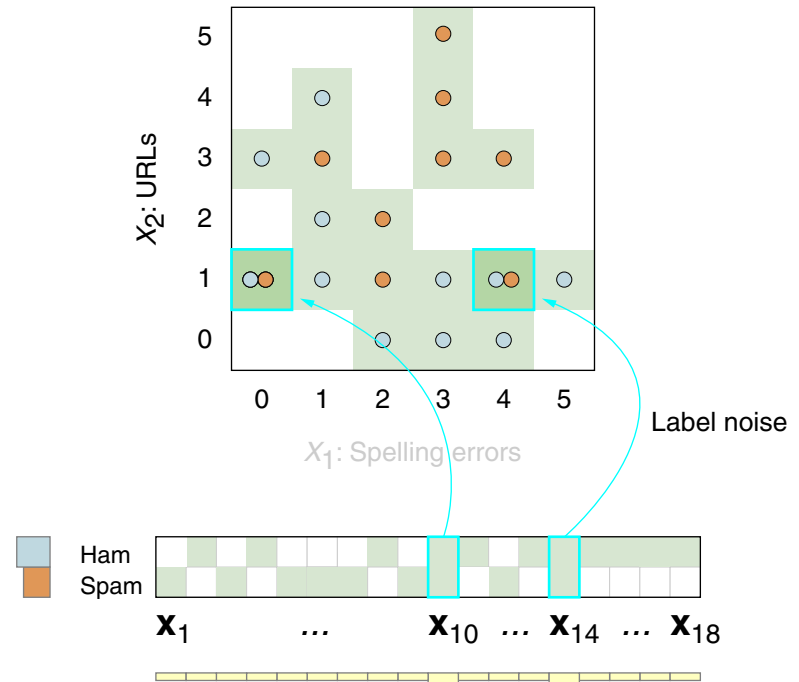
- Approximate D to compute a classifier for feature vectors \mathbf{x} , $\mathbf{x} \notin D$.

Frequentist versus Subjectivist

Logistic Regression versus Naive Bayes: Example (continued)

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:

- Approximate D to compute a classifier for feature vectors x , $x \notin D$.

Frequentist versus Subjectivist

Logistic Regression versus Naive Bayes: Conditional Class Probabilities

Logistic regression:

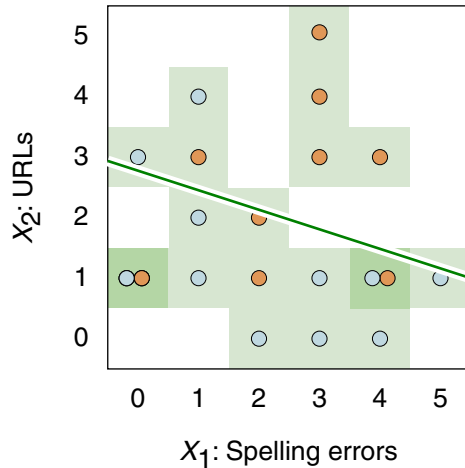


□ Distribution of D .

Frequentist versus Subjectivist

Logistic Regression versus Naive Bayes: Conditional Class Probabilities (continued)

Logistic regression:

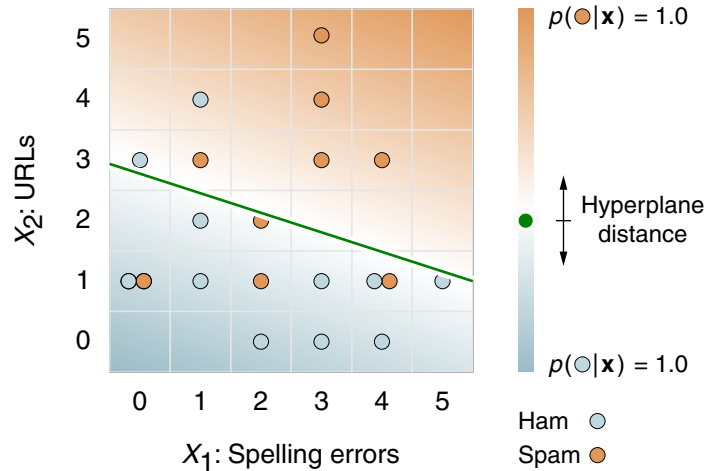


□ Hyperplane $\langle \mathbf{w}_{\text{ML}}, \mathbf{x} \rangle = 0$.

Frequentist versus Subjectivist

Logistic Regression versus Naive Bayes: Conditional Class Probabilities (continued)

Logistic regression:

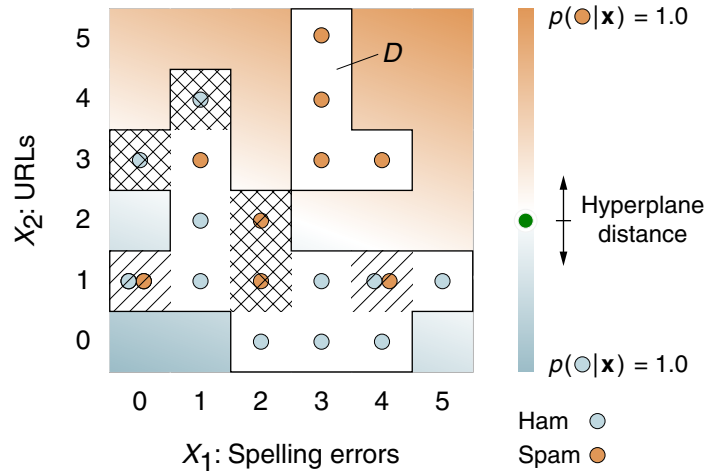


- Conditional class probabilities computed with \mathbf{w}_{ML} , the ML estimate for \mathbf{w} given D .

Frequentist versus Subjectivist

Logistic Regression versus Naive Bayes: Conditional Class Probabilities (continued)

Logistic regression:

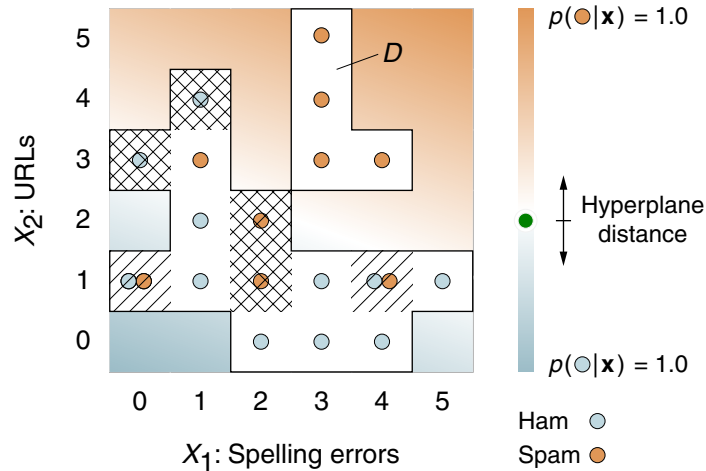


□ Training error.

Frequentist versus Subjectivist

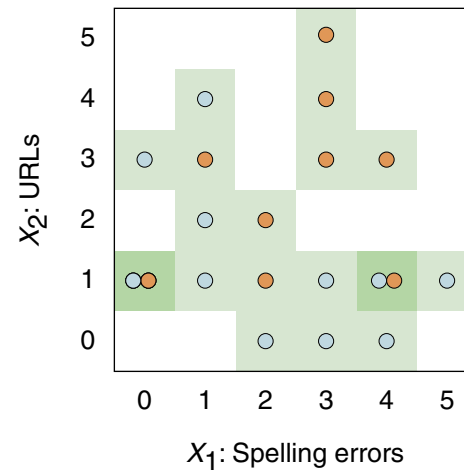
Logistic Regression versus Naive Bayes: Conditional Class Probabilities (continued)

Logistic regression:



□ Training error.

Naive Bayes:

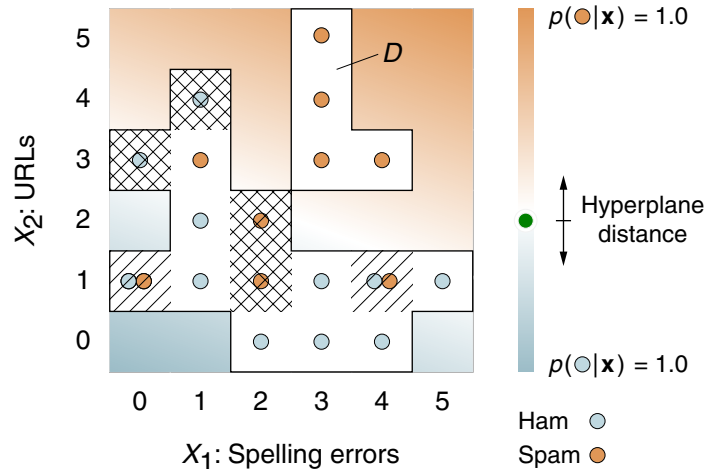


□ Distribution of D .

Frequentist versus Subjectivist

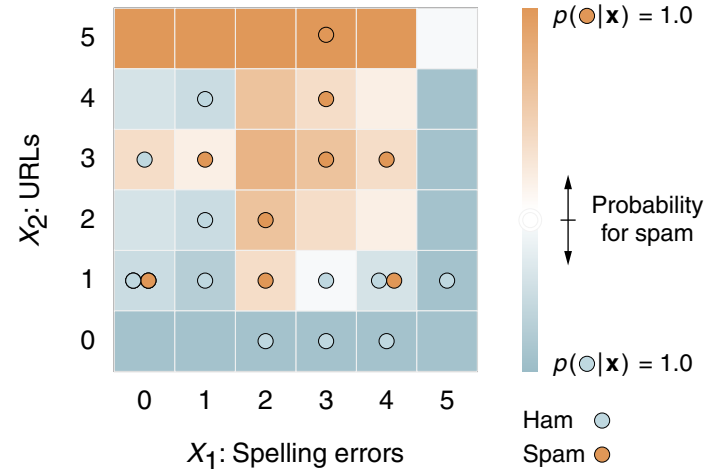
Logistic Regression versus Naive Bayes: Conditional Class Probabilities (continued)

Logistic regression:



- Training error.

Naive Bayes:

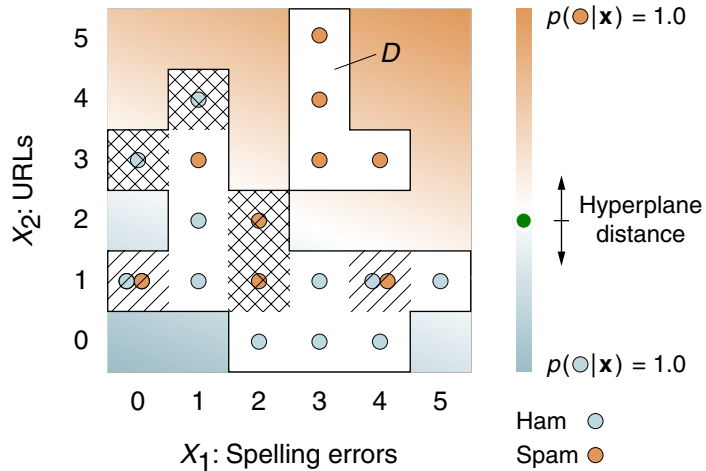


- Conditional class probabilities computed for the respective MAP class, using $p(c)$ estimates from D .

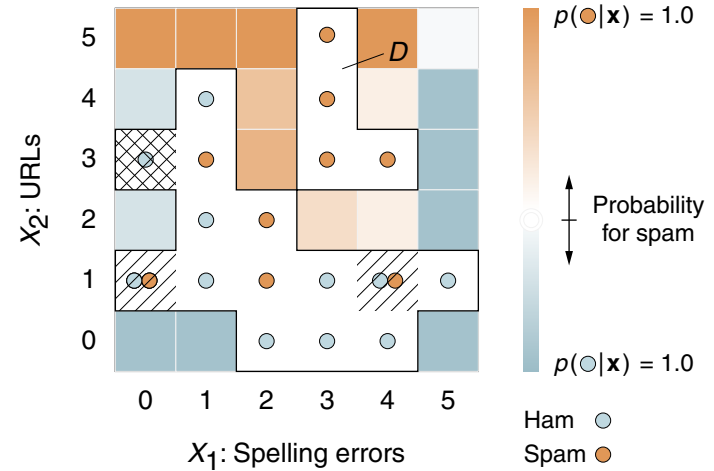
Frequentist versus Subjectivist

Logistic Regression versus Naive Bayes: Conditional Class Probabilities (continued)

Logistic regression:



Naive Bayes:



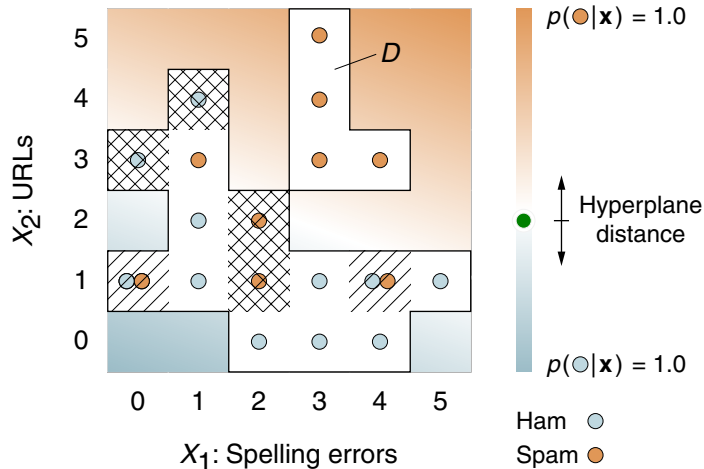
□ Training error.

□ Training error.

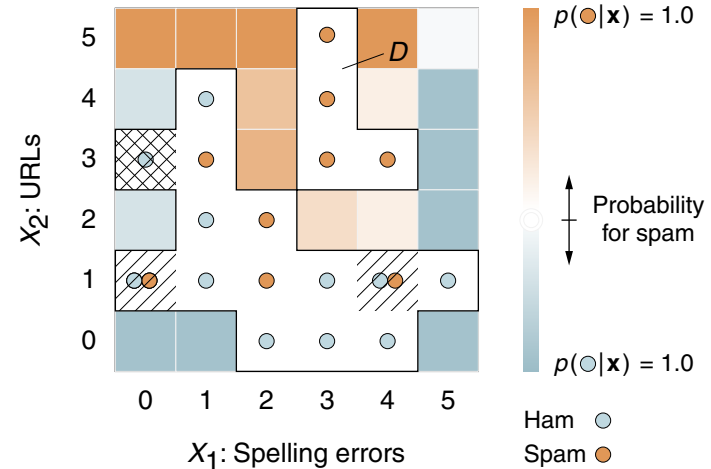
Frequentist versus Subjectivist

Logistic Regression versus Naive Bayes: Conditional Class Probabilities (continued)

Logistic regression:



Naive Bayes:



- ❑ Computation of a hyperplane.
- ❑ Approach: minimization of accumulated “misclassification distances” for examples in D .
- ❑ Discriminative and probabilistic.

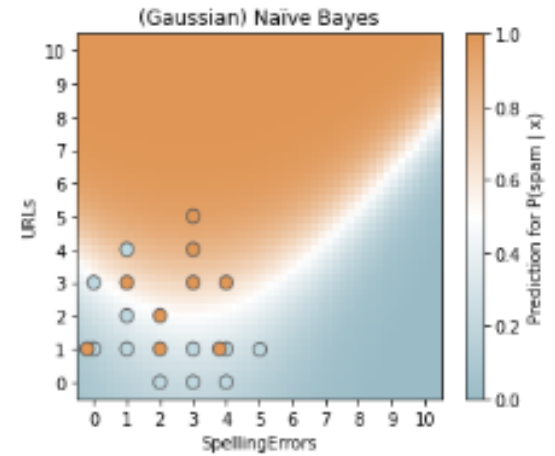
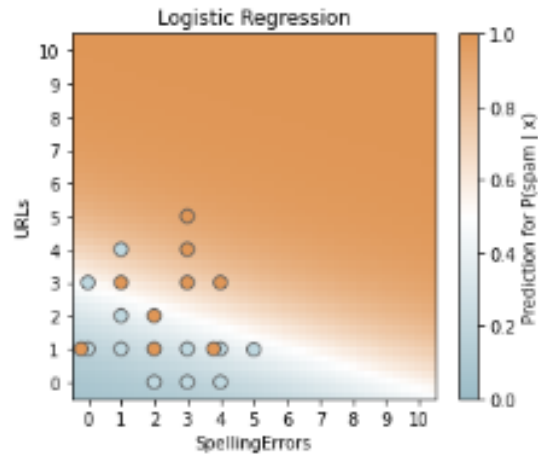
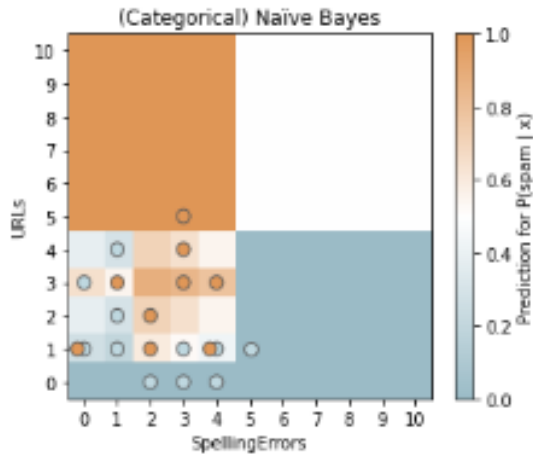
- ❑ Computation of a probability distribution.
- ❑ Basis: class-conditional feature and class frequencies in D .
- ❑ Generative (implies probabilistic).

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 \mathbb{R}^2 . (Whether that makes sense is another matter.)
 - 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} 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

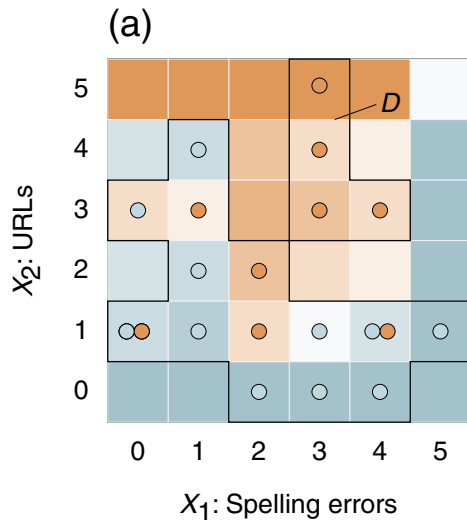


\leadsto *BOARD*

Frequentist versus Subjectivist

Naive Bayes: Prior Probability Models

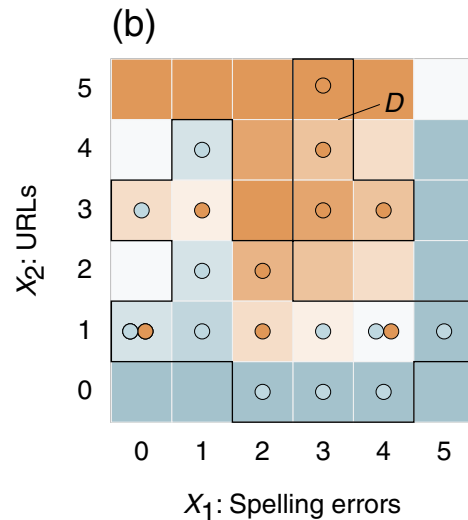
Comparison of the conditional class probability function, $p(c \mid \mathbf{x})$, under Naive Bayes for three different prior probability models (= assessments of class priors), $p(c)$.



$p(c)$ estimates from D

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

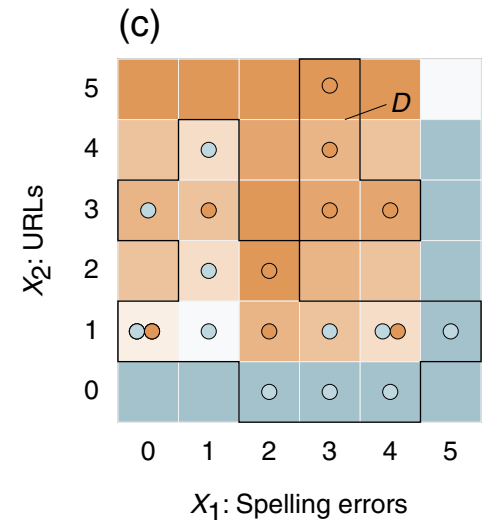
$$P_a(C=\text{Ham}) = \hat{p}(\text{Ham}) = 0.55$$



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

\leadsto *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”, $H=h$, and a “data event”, $\mathbf{D}=D$,

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

Frequentist versus Subjectivist

Advanced Bayesian Decision Making (continued)

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”, $H=h$, and a “data event”, $\mathbf{D}=D$,

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

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 (continued)

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”, $H=h$, and a “data event”, $\mathbf{D}=D$,

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

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 (continued)

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”, $H=h$, and a “data event”, $\mathbf{D}=D$,

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

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 (continued)

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”, $H=h$, and a “data event”, $\mathbf{D}=D$,

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

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 Bayes method 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., \mathbf{D} is a random vector whose domain is the feature space \mathbf{X} .
- When using the Bayes method for an outcome-only setting, then $p(D)$, $p(D) := P(\mathbf{D}=D)$, is the probability of the data $D = \{y_1, \dots, y_n\}$ or $D = \{c_1, \dots, c_n\}$. I.e., \mathbf{D} is a random vector whose domain is \mathbb{R}^n or C^n , where C is the set of possible classes or class labels.
- $p(h) := P(H=h)$ (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.
- Recall that $p(\cdot)$ is defined via $P(\cdot)$ and that the two notations can be used interchangeably, arguing about realizations of random variables and events respectively.