BAYESIAN LEARNING

Fernando Pérez Cruz

25 de October 2004

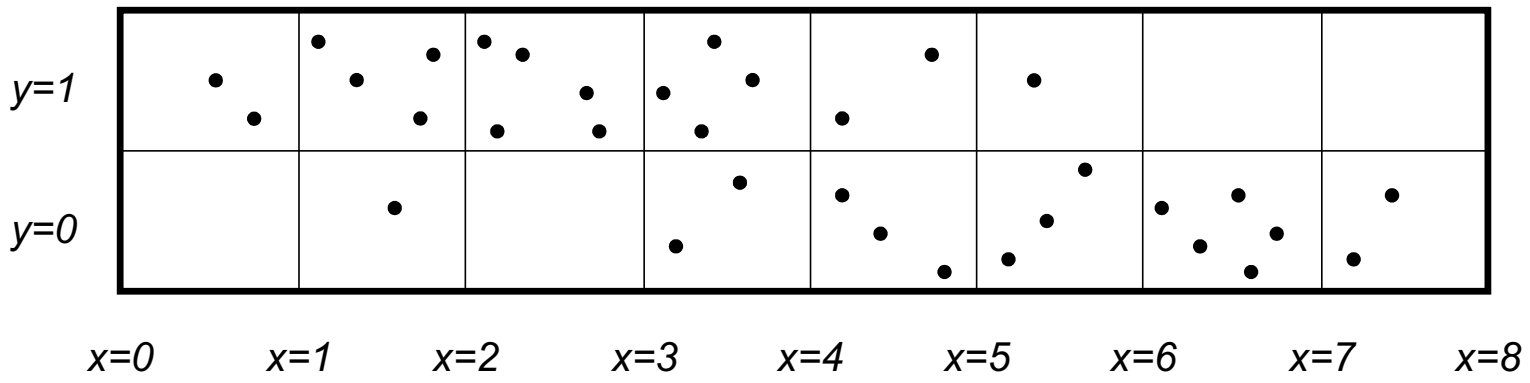
INDEX

BAYES RULE

- Bayes Rule:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

- p(1<x<2 | y=0)?



MACHINE LEARNING USING BAYES RULE

- Regression estimation:

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots (\mathbf{x}_n, y_n)\} \qquad \mathbf{x}_i \in \mathbb{R}^d, \quad y \in \mathbb{R}$$

- We have 4 different hypothesis: $h_1(\mathbf{x})$, $h_2(\mathbf{x})$, $h_3(\mathbf{x})$ and $h_4(\mathbf{x})$.
- Compute the output for $\mathbf{x}^*$ new input vector.
- We use Bayes Rule to obtain the posterior for each hypothesis:

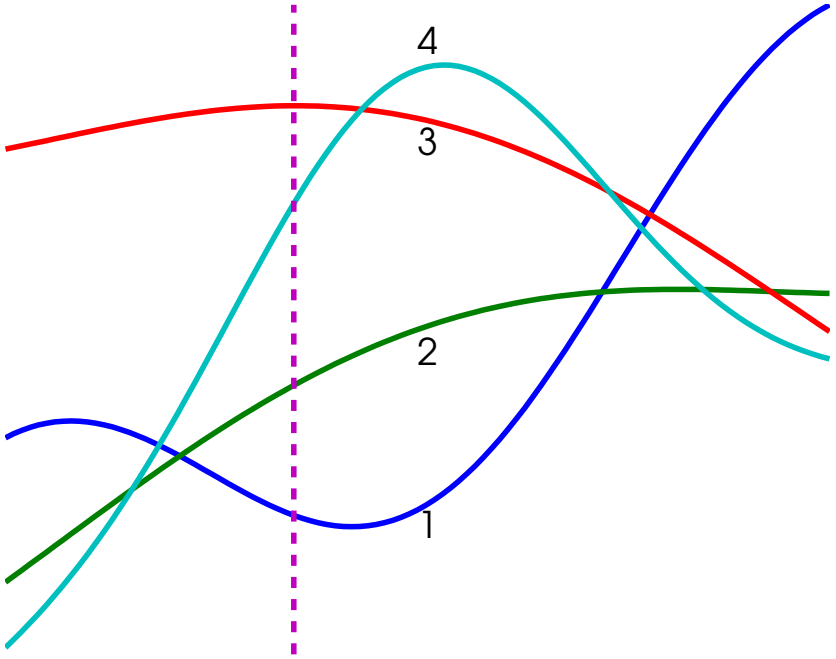
$$p(h_i|\mathcal{D}) = \frac{p(\mathcal{D}|h_i)p(h_i)}{p(\mathcal{D})} \qquad \forall i = 1, \ldots, 4.$$

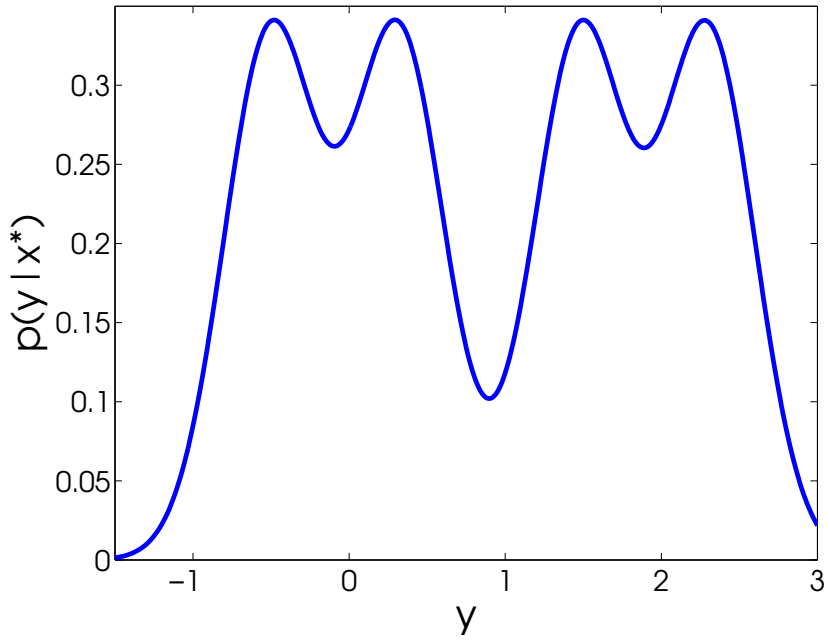
- We predict the output as a weighted sum:

$$p(y^*|\mathbf{x}^*, \mathcal{D}) = \sum_{i=1}^{4} p(y^*|h_i, \mathbf{x}^*)p(h_i|\mathcal{D})$$

EXAMPLE

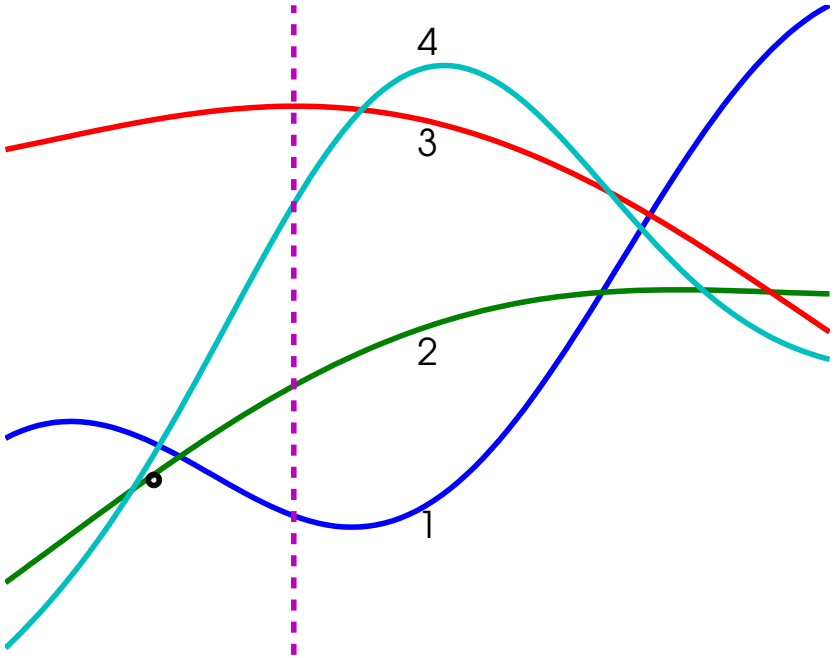
- Four Hypothesis. Predict the output (Gaussian Noise $\sigma = 0.3$).

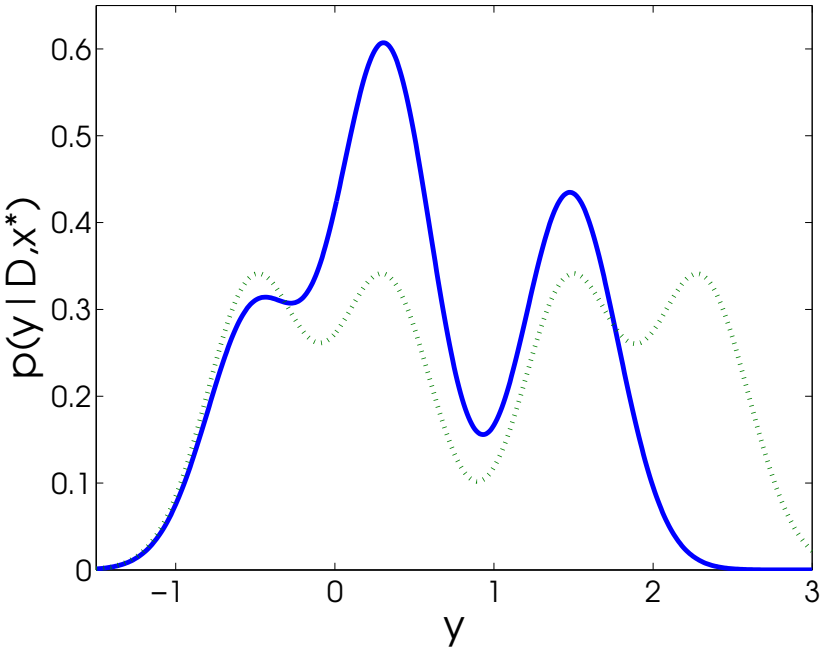




First Data Point

- $x_1 = 0.18, y_1 = -0.29.$



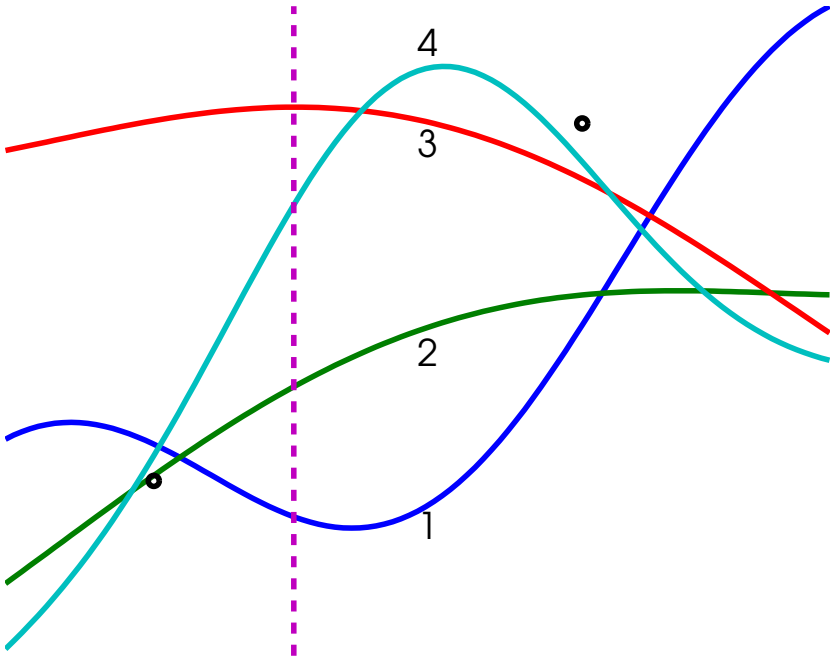


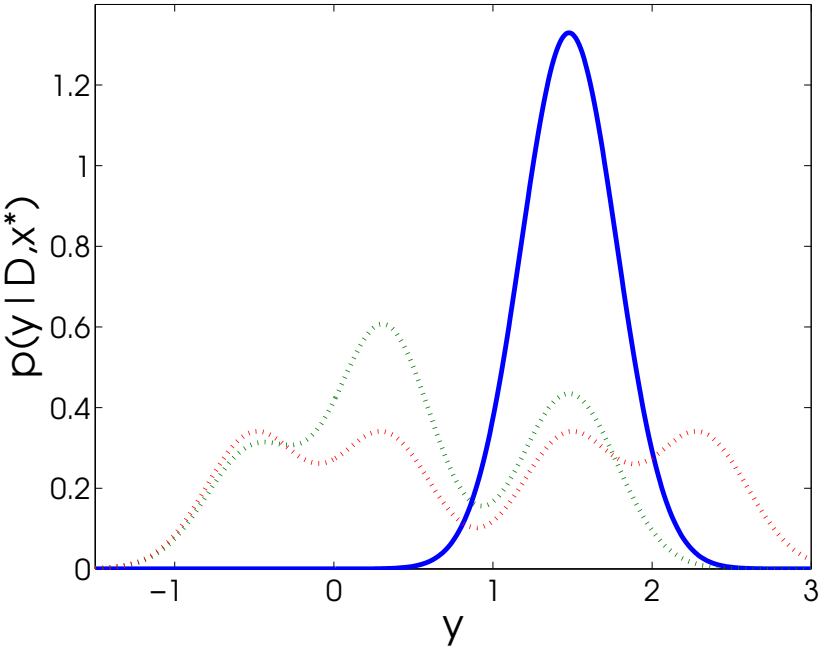
- Posterior = [0.22, 0.45, 0.00, 0.33]

$P(h_3|D) \sim 4 \cdot 10^{-31}.$

Second Data Point

- $x_2 = 0.70, y_2 = 2.19.$





- Posterior = [0.00, 0.00, 0.00, 1.00]

$P(h_2|D) \sim 5 \cdot 10^{-7}.$

MAXIMUM A POSTERIORI

- Computing the sum over all the hypothesis might not be possible.
- Use the most probable hypothesis:
- Maximum a posteriori (MAP):

$$\widehat{h} = \underset{h_i}{\text{argmax}} \frac{p(\mathcal{D}|h_i)p(h_i)}{p(\mathcal{D})} = \underset{h_i}{\text{argmax}}\, p(\mathcal{D}|h_i)p(h_i)$$

- We predict the output as:

$$p(y^*|\widehat{h}, \mathbf{x}^*)$$

Maximum Likelihood

- In the case we have not prior information to prefer any solution.
- Flat prior
- Maximum Likelihood (ML):

$$\widehat{h} = \operatorname{argmax}_{h_i} p(\mathcal{D}|h_i)$$

- We predict the output as:

$$p(y^*|\widehat{h}, \mathbf{x}^*)$$

HYPERPARAMETER ESTIMATION

- Most cases we will have infinite number of hypothesis:

$$h(x) = \sum_{i=1}^{P} w_j x^P + w_0 \qquad\qquad w_j \in \mathbb{R}$$

- For Gaussian noise, the likelihood of each model will be:

$$P(\mathcal{D}|\mathbf{w}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - h(x_i))^2}{2\sigma^2}\right)$$

- For Gaussian Prior:

$$P(\mathbf{w}) = \frac{1}{\sqrt{|2\pi\Sigma_{\mathbf{w}}|}} \exp\left(-\frac{1}{2}\mathbf{w}^\top \Sigma_{\mathbf{w}}^{-1} \mathbf{w}\right)$$

- How do we set P, σ or $\Sigma_{\mathbf{w}} (= \sigma_{\mathbf{w}} \mathbf{I})$?

HYPERPARAMETER ESTIMATION II

- In this case, we have that:

$$p(y^*|\mathbf{x}^*, \mathcal{D}, \theta) = \int p(y^*|\mathbf{w}, \mathbf{x}^*, \theta) p(\mathbf{w}|\mathcal{D}, \theta) d\mathbf{w}$$

  where $\theta = (P, \sigma, \sigma_w, \ldots)$.

- We can integrate θ out as we did with the model:

$$p(y^*|\mathbf{x}^*, \mathcal{D}) = \int p(y^*|\mathbf{x}^*, \mathcal{D}, \theta) p(\theta|\mathcal{D}) d\theta$$

- And use Bayes Rule to obtain the posterior of the hyperparameters:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) p(\theta)}{p(\mathcal{D})}$$

- Or use the Maximum Likelihood solution: $p(y^*|\mathbf{x}^*, \mathcal{D}) = p(y^*|\mathbf{x}^*, \mathcal{D}, \theta_{ML})$

TEXT CATEGORIZATION

- We have a document: web page, newspaper article, ...
- We want to assign it to a category: politics, economics, sport, social ...
- Preprocessing:
  - Select keywords.
  - Count how many times they do appear in the text to form the input vector.
- Select the category by Maximum a Posteriory:

$$\widehat{y} = \underset{y_i}{\text{argmax}}\, p(y_i|x_1, x_2, \ldots, x_d) = \underset{y_i}{\text{argmax}}\, \frac{p(x_1, x_2, \ldots, x_d|y_i)p(y_i)}{p(x_1, x_2, \ldots, x_d)}$$

NAÏVE BAYES

- Suppose d=100 and that x_j can only take 0 or 1.
- If we want it only one example per possible input, we will need $2^{100} \simeq 1.3 \cdot 10^{30}$ documents.
- Naïve Bayes: Suppose that given y_i the input components are independent.

$$p(x_1, x_2, \ldots, x_d|y_i)p(y_i) = \prod_{j=1}^{d} p(x_j|y_i)$$

- With only 2 examples we will have covered the hole space.
- Naïve Bayes is not a Bayesian approach.

EXAMPLE: NAÏVE BAYES

- Predict if we are going to play tennis today (sunny, cool, high, strong):

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

SOLUTION

- We cannot the answer directly from $p(\text{sunny}, \text{cool}, \text{high}, \text{strong}|\text{yes})$ and $p(\text{sunny}, \text{cool}, \text{high}, \text{strong}|\text{no})$, because they are both zero.
- We can use Naïve Bayes:
  - Prior Probabilities: $p(y = \text{no}) = 5/14$ and $p(y = \text{yes}) = 9/14$.
  - Likelihood Yes:

    $p(\text{sunny}, \text{cool}, \text{high}, \text{strong}|\text{yes}) = p(\text{sunny}|\text{yes})p(\text{cool}|\text{yes})p(\text{high}|\text{yes})p(\text{strong}|\text{yes})$
  - Likelihood No:

    $p(\text{sunny}, \text{cool}, \text{high}, \text{strong}|\text{no}) = p(\text{sunny}|\text{no})p(\text{cool}|\text{no})p(\text{high}|\text{no})p(\text{strong}|\text{no})$
- We can compute this probabilities from the data table.

PROBABILITIES

$p(x_j \mid y_i)$	Sunny	Cool	High	Strong
no	3/5	1/5	4/5	3/5
yes	2/9	3/9	3/9	3/9

$$p(\text{sunny}, \text{cool}, \text{high}, \text{strong} \mid \text{no}) = \frac{1}{p(\text{sunny}, \text{cool}, \text{high}, \text{strong})} \times \frac{3 \times 1 \times 4 \times 3}{5 \times 5 \times 5 \times 5} \times \frac{5}{14}$$
$$= \frac{0.0206}{p(\text{sunny}, \text{cool}, \text{high}, \text{strong})}$$

$$p(\text{sunny}, \text{cool}, \text{high}, \text{strong} \mid \text{yes}) = \frac{1}{p(\text{sunny}, \text{cool}, \text{high}, \text{strong})} \times \frac{2 \times 3 \times 3 \times 3}{9 \times 9 \times 9 \times 9} \times \frac{9}{14}$$
$$= \frac{0.0053}{p(\text{sunny}, \text{cool}, \text{high}, \text{strong})}$$

$p(\text{sunny}, \text{cool}, \text{high}, \text{strong} \mid \text{no}) = 0.795$ $\quad$ $p(\text{sunny}, \text{cool}, \text{high}, \text{strong} \mid \text{yes}) = 0.205$