Machine-learning Crash Course

$$\overbrace{p(\mathbf{w}|\mathcal{D})}^{\text{posterior}} = \frac{\overbrace{p(\mathcal{D}|\mathbf{w})}^{\text{likelihood}}\,\overbrace{p(\mathbf{w})}^{\text{prior}}}{p(\mathcal{D})}$$

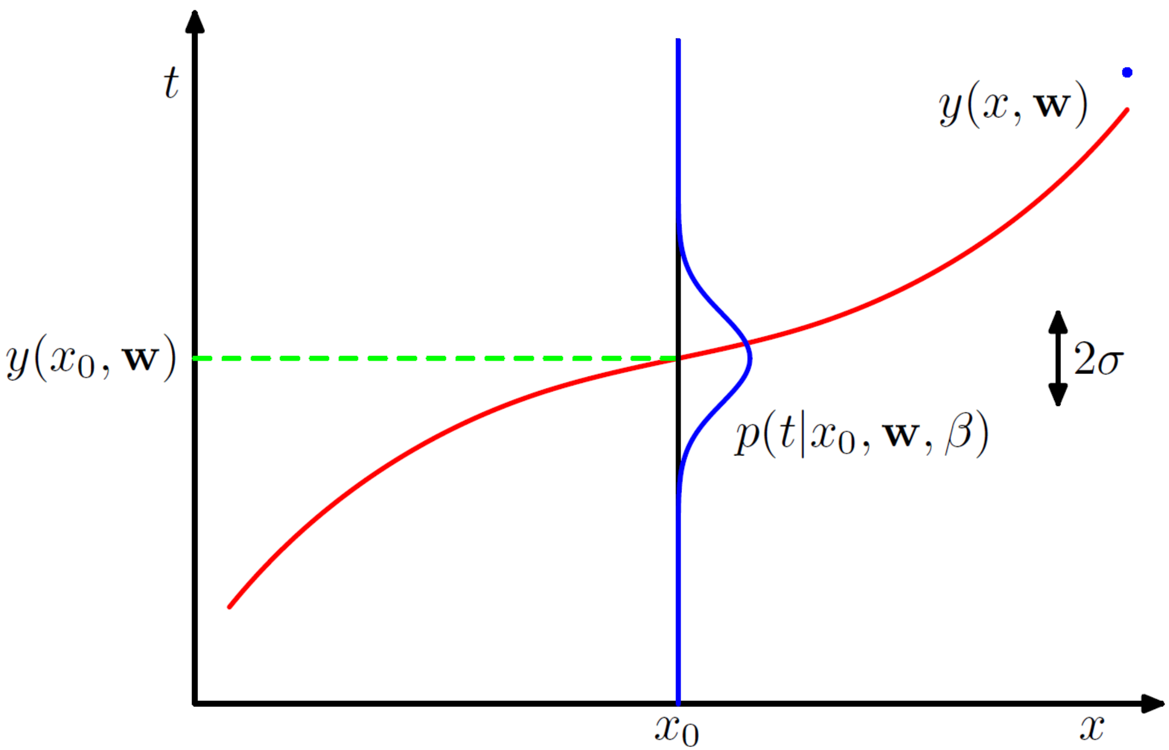
M*aximum likelihood (frequentist approach)*:

- **w** is set to the value that maximizes the likelihood function *p*(*D*/**w**).
- This corresponds to choosing the value of **w** for which the probability of the observed data set is maximized.
- In the machine learning literature, the negative log of the likelihood function is called an *error function*.

A more probabilistic approach to curve fitting:

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right)$$

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M$$



$$\text{training data } \{\mathbf{x}, \mathbf{t}\}$$

Likelihood function:

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n | y(x_n, \mathbf{w}), \beta^{-1}\right)$$

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \cancel{\frac{N}{2} \ln \beta} - \cancel{\frac{N}{2} \ln(2\pi)}$$

The sum-of-squares error function arises as a consequence of **maximizing likelihood under the assumption of a Gaussian noise distribution.**

$$\boxed{p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t | y(x, \mathbf{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}\right)}$$

predictive distribution that gives the probability distribution over *t*, rather than simply a point estimate

Maximum a Posteriori (MAP):

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

hyperparameter

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

$$\boxed{\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}}$$

maximizing the posterior distribution is equivalent to minimizing the regularized sum-of-squares error function

Maximum Likelihood and kullback-Leibler Divergence

Let $p_{\text{model}}(\mathbf{x};\boldsymbol{\theta})$ be a parametric family of probability distributions over the same space indexed by $\boldsymbol{\theta}$. In other words, $p_{\text{model}}(\boldsymbol{x};\boldsymbol{\theta})$ maps any configuration $\boldsymbol{x}$ to a real number estimating the true probability $p_{\text{data}}(\boldsymbol{x})$.

$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} p_{\text{model}}(\mathbb{X};\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log p_{\text{model}}(\boldsymbol{x}^{(i)};\boldsymbol{\theta})$$

$$\downarrow$$

$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x}\sim\hat{p}_{\text{data}}} \log p_{\text{model}}(\boldsymbol{x};\boldsymbol{\theta})$$

$$\wr$$

$$D_{\text{KL}}\left(\hat{p}_{\text{data}} \| p_{\text{model}}\right) = \mathbb{E}_{\mathbf{x}\sim\hat{p}_{\text{data}}}\left[\log \hat{p}_{\text{data}}(\boldsymbol{x}) - \log p_{\text{model}}(\boldsymbol{x})\right]$$

Goodfellow, Bengio, Courville - Deep Learning 2016

Conditional Maximum Likelihood

This is actually the most common situation because it forms the basis for most supervised learning.

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{Y} \mid \boldsymbol{X}; \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log P(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

Example: probabilistic curve fitting

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n | y(x_n, \mathbf{w}), \beta^{-1}\right)$$

Example, Maximum likelihood and the Gaussian parameters:

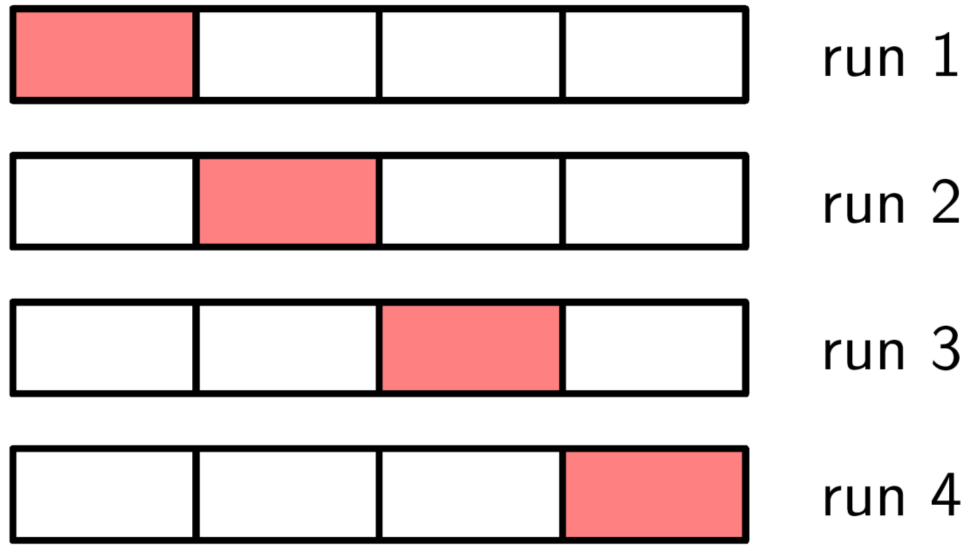
$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu, \sigma^2\right)$$

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \longrightarrow \mathbb{E}[\mu_{\mathrm{ML}}] = \mu$$

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2 \longrightarrow \mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2$$

Model selection

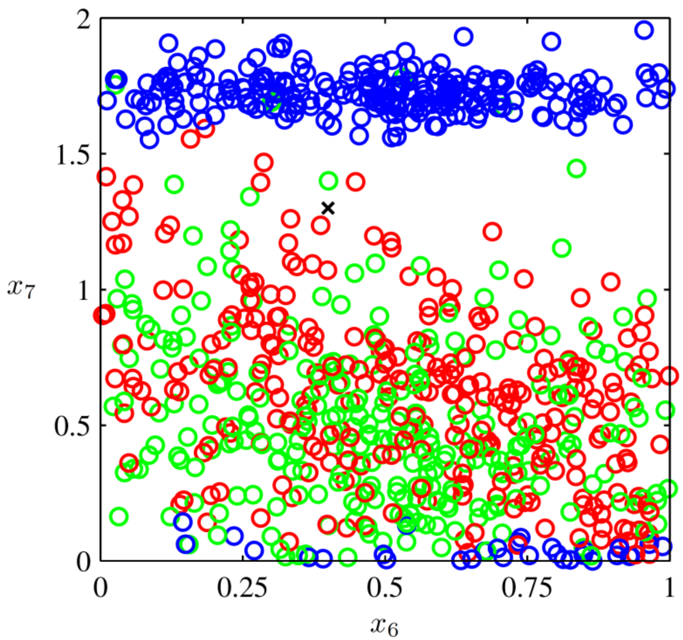
S-fold cross-validation (S=1 → leave-one-out)

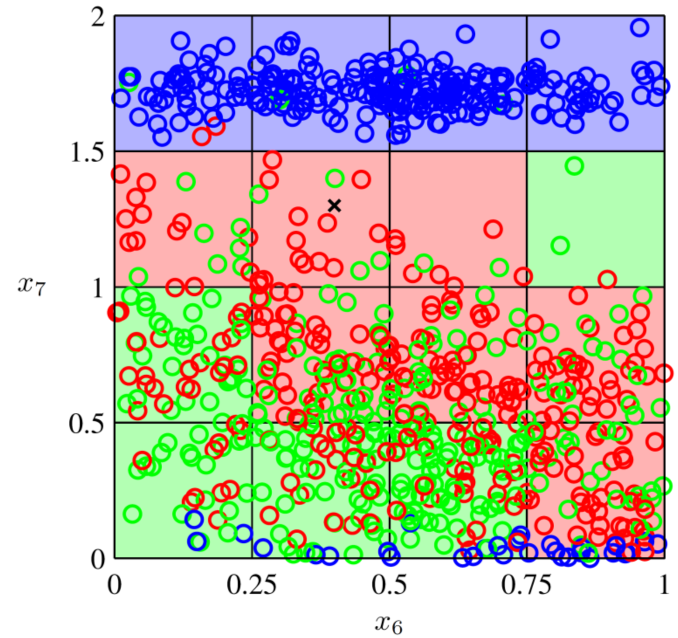


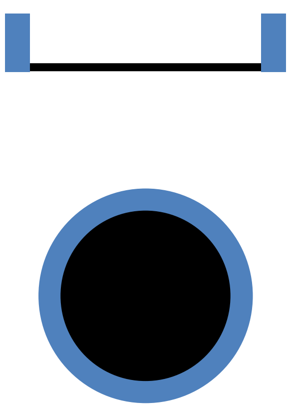
Akaike Information Criterion (AIC):

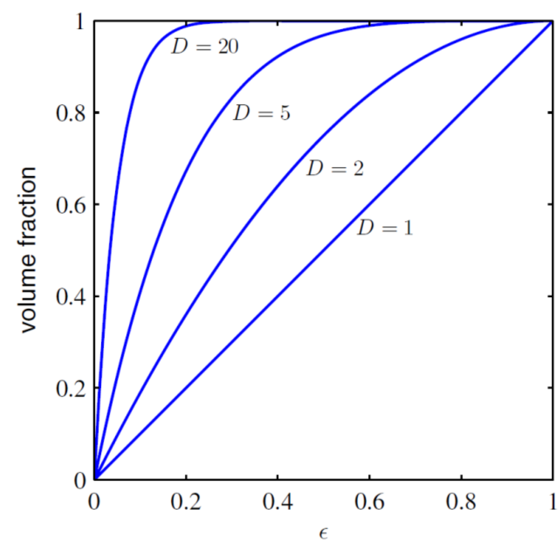
$$\ln p(\mathcal{D}|\mathbf{w}_{\mathrm{ML}}) - M$$

The curse of dimensionality

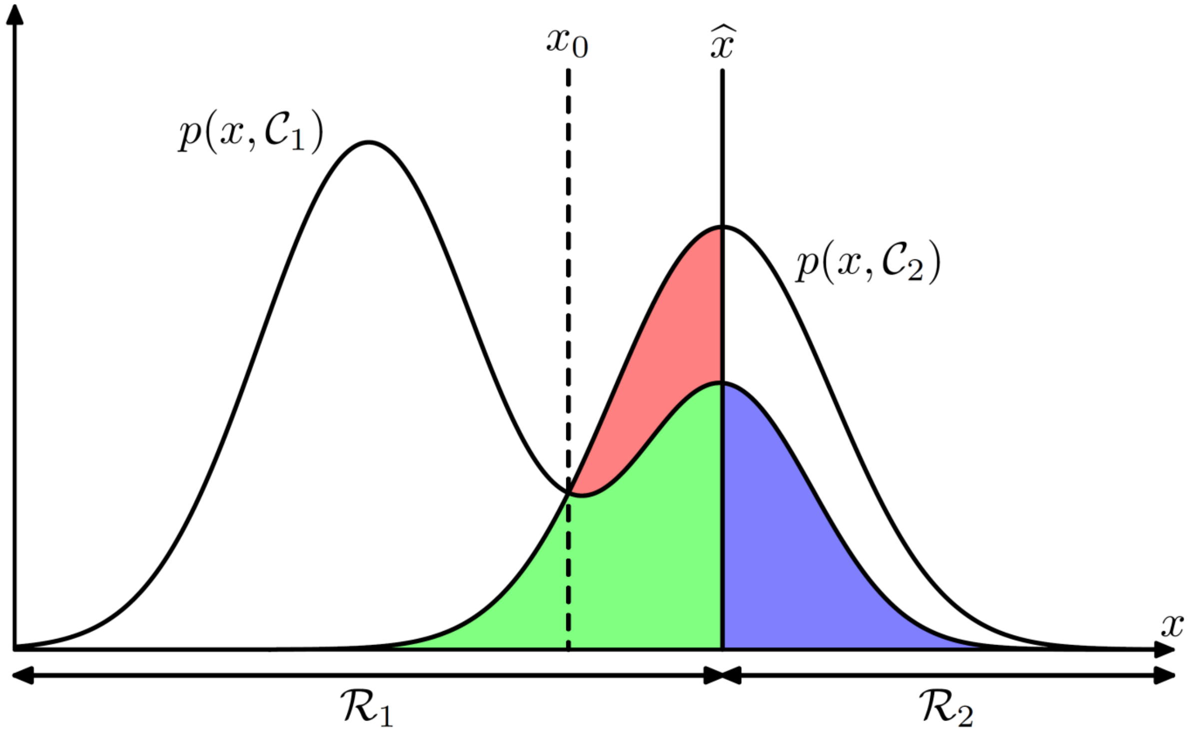








Decision theory



$$
\begin{aligned}
p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\
&= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}.
\end{aligned}
$$

$$= p(\mathcal{C}_k|\mathbf{x})\cancel{p(\mathbf{x})}$$

The minimum probability of making a mistake is obtained if each value of **x is assigned to the class for which the posterior probability *p*(*Ck*/x) is largest**

Minimizing the expected loss

$$\begin{array}{c c} & \begin{array}{cc} \text{cancer} & \text{normal} \end{array} \\ \begin{array}{c} \text{cancer} \\ \text{normal} \end{array} & \left(\begin{array}{cc} 0 & 1000 \\ 1 & 0 \end{array} \right) \end{array}$$

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) \, \mathrm{d}\mathbf{x}$$

Ways of solving a decision problem

1. First solve the inference problem of determining the class-conditional densities $p(\mathbf{x}/Ck)$ and the prior class probabilities $p(Ck)$. Then use Bayes' theorem to compute the posterior:

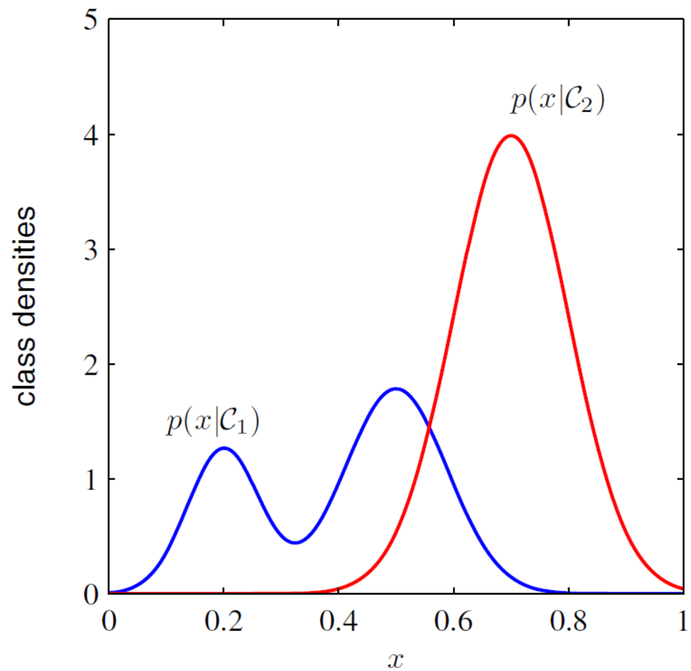
$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

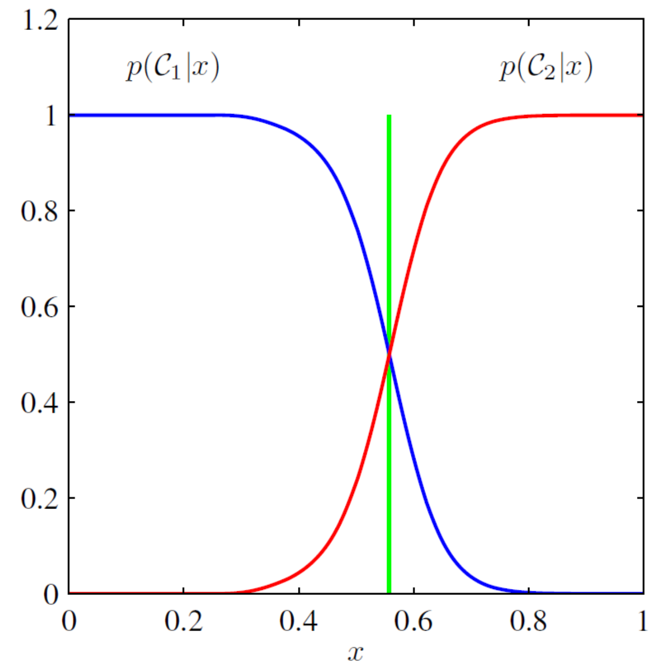
We can also estimate $p(\mathbf{x},Ck)$ via **generative models** (Gaussian mixture model or Generative adversarial Nets). This is usually a difficult problem.

2. **Determine the posterior class probabilities p(Ck|x)**, and then subsequently use decision theory to assign each new **x** to one of the classes. Approaches that model the posterior probabilities directly are called **discriminative models** (logistic regression, GLM).

3. Find a function $f(\mathbf{x})$, called a **discriminant function**, which maps each input **x** directly onto a class label. Probabilities play no role (LDA?).

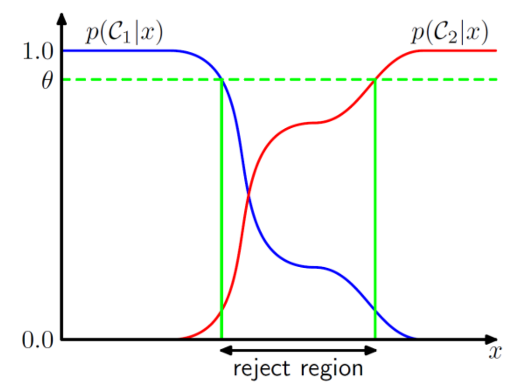
Pros and cons





Having p(Ck|x):

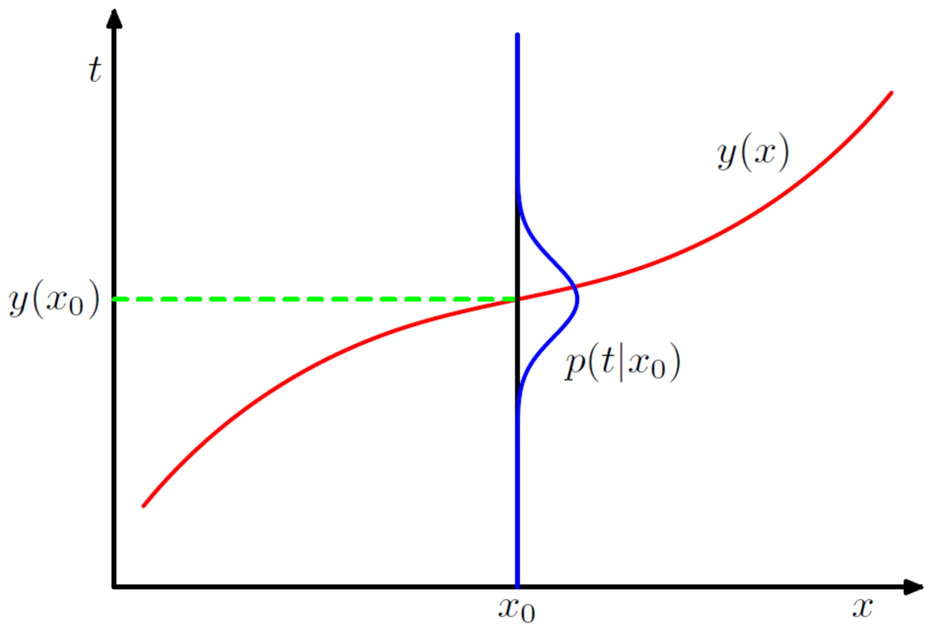
- Risk minimization: updating the loss matrix.
- Reject region.
- Compensating for class priors.
- Combining models:



$$
\left.\begin{aligned}
p(\mathcal{C}_k|\mathbf{x}_\mathrm{I}, \mathbf{x}_\mathrm{B}) &\propto p(\mathbf{x}_\mathrm{I}, \mathbf{x}_\mathrm{B}|\mathcal{C}_k)p(\mathcal{C}_k) \\
&\propto p(\mathbf{x}_\mathrm{I}|\mathcal{C}_k)p(\mathbf{x}_\mathrm{B}|\mathcal{C}_k)p(\mathcal{C}_k) \\
&\propto \frac{p(\mathcal{C}_k|\mathbf{x}_\mathrm{I})p(\mathcal{C}_k|\mathbf{x}_\mathrm{B})}{p(\mathcal{C}_k)}
\end{aligned}\right\} \text{Naïve bayes}
$$

Assuming conditional independency

Decision theory and regression



$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

The error component we try to minimize

the variance of the distribution of t, averaged over x (intrinsic variability)