Comitês de Máquinas

1. Prelúdio e Analogia

- Se tivermos de tomar uma decisão assaz importante, provavelmente pensaremos em consultar pessoas qualificadas que possam nos auxiliar. Nesse caso, seria prudente, por certo, ter conselheiros competentes. Ademais, seria importante que as opiniões desses conselheiros trouxessem diferentes perspectivas sobre o tema em análise. Afinal, se todos pensassem da mesma forma sobre tudo, bastaria contar com um deles.
- Essa ideia pode ser estendida ao aprendizado de máquina: diferentes máquinas podem resolver certo problema de maneira distinta, o que torna pertinente a ideia de combinar essas máquinas em um **comitê** [Tresp, 2001].

- A diversidade de soluções pode tornar o comitê mais robusto em termos de generalização. Isso se explica porque a diversificação de máquinas é uma "diversificação de generalizações" [Coelho et al., 2016].
- O processo é análogo ao de construção de um portfólio de ações: diversifica-se a composição desse portfólio para diminuir o risco total [Markowitz, 1952].

2. Comitês de Máquinas

- Um **comitê de máquinas** é uma estratégia baseada na combinação de diferentes estruturas. Há dois tipos principais de comitês: estáticos e dinâmicos [Coelho et al., 2016]. Os **comitês estáticos** possuem um mecanismo de combinação de respostas que não depende da informação de entrada. Por outro lado, um **comitê dinâmico** utiliza a entrada para integrar a resposta dos especialistas

(como nas misturas de especialistas). As Figs. 1 e 2 ilustram os dois paradigmas.

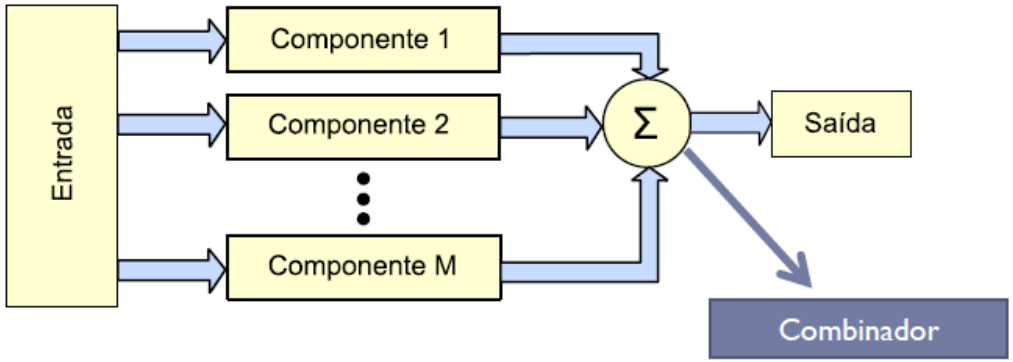

Figura 1. Comitê Estático (de [Coelho et al., 2016]).

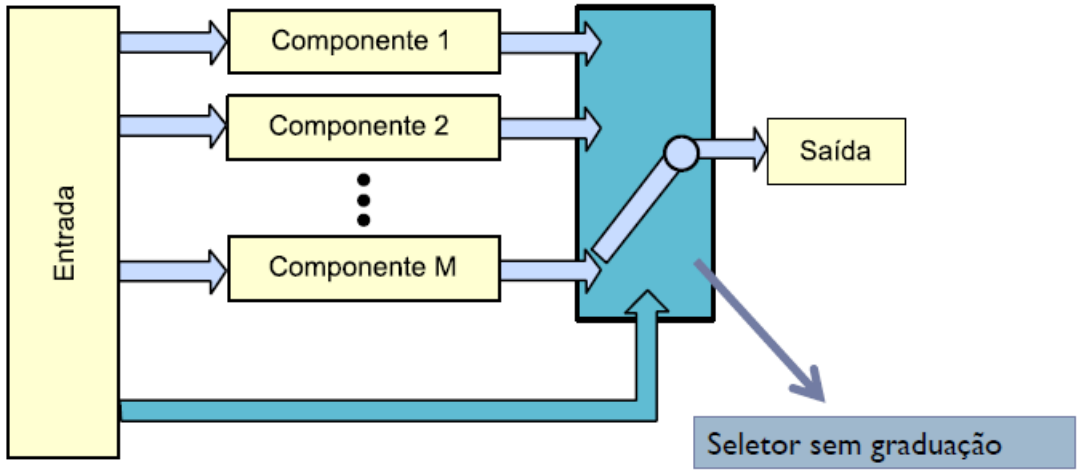

Figura 2. Comitê Dinâmico (de [Coelho et al., 2016]).

- Discutiremos primeiramente a abordagem estática paradigmática, a de *ensembles*.

2.1. Ensembles: Definição, Bagging e Boosting

- Num *ensemble*, M máquinas são treinadas a partir dos dados disponíveis e, a partir de algum tipo de combinação de suas saídas, gera-se a resposta global. Vale a pena revisitar a Fig. 1.
- Conforme discutimos anteriormente, a ideia de combinar máquinas tem por base a ideia de que "diversificar perspectivas" pode trazer uma melhor generalização [Bishop, 2006].
- Para que entendamos o porquê disso, consideremos um problema de regressão em que se deseja aproximar uma função ideal $h(\mathbf{x})$ a partir de um conjunto de dados. Dispomos de um ensemble com M modelos de regressão, que geram as

estimativas $y_1(\mathbf{x}), y_2(\mathbf{x}), \ldots, y_M(\mathbf{x})$. Consideraremos que a saída do comitê será a média aritmética dessas estimativas[1] [Bishop, 2006]:

$$y_C(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} y_m(\mathbf{x})$$

- Consideraremos que a saída do m-ésimo modelo é interpretada como:

$$y_m(\mathbf{x}) = h(\mathbf{x}) + e_m(\mathbf{x})$$

Isso significa que $e_m(\mathbf{x})$ é o erro associado ao m-ésimo modelo. O m-ésimo erro quadrático médio será:

$$\text{EQM}_m = E[{e_m}^2(\mathbf{x})] = E\left[\left(y_m(\mathbf{x}) - h(\mathbf{x})\right)^2\right]$$

- O erro médio gerado pelos modelos *individualmente* é:

[1] Vale mencionar que, em problemas de classificação, a agregação por voto majoritário é uma opção usual. Outras opções são discutidas em [Coelho et al., 2016].

$$\text{EQM}_{\text{médio}} = \frac{1}{M}\sum_{m=1}^{M} E[{e_m}^2(\mathbf{x})] = \frac{1}{M}\sum_{m=1}^{M} EQM_m$$

- Analisemos agora o erro associado ao comitê. Temos:

$$\text{EQM}_C = E\left[\left(\frac{1}{M}\sum_{m=1}^{M} y_m(\mathbf{x}) - h(\mathbf{x})\right)^2\right] = E\left[\left(\frac{1}{M}\sum_{m=1}^{M} e_m(\mathbf{x})\right)^2\right]$$

- Se supusermos que os erros $e_m(\mathbf{x})$ têm média nula e são descorrelacionados, obtém-se que:

$$\text{EQM}_C = \frac{1}{M}\text{EQM}_{\text{médio}}$$

Isso quer dizer que o *ensemble* terá um erro quadrático médio menor que a média dos erros individuais ($\text{EQM}_{\text{médio}}$), o que significa que a combinação seria proveitosa. No entanto, note que supusemos uma diversificação idealizada, pois consideramos que os erros eram *descorrelacionados*.

- Um caminho para buscar aproximar essa condição é realizar um procedimento de *bootstrap* para gerar os conjuntos de dados das máquinas individuais. Nesse procedimento, se houver N dados no conjunto de treinamento, cada conjunto será composto de N amostras obtidas com reposição [Bishop, 2006]. Isso significa que pode haver dados repetidos e, por consequência, dados ausentes em cada conjunto.
- A estratégia acima, de agregação com *bootstrap*, é conhecida como *bootstrap aggregation* ou *bagging* [Bishop, 2006].
- Outra abordagem clássica é a de *boosting*. Nesse caso, adota-se um esquema de treinamento sequencial, ou seja, as máquinas são treinadas em sequência. O treinamento de cada máquina se baseia num *dataset* em que os dados são ponderados de acordo com o desempenho das "máquinas anteriores" [Bishop, 2006].

- Consideremos um exemplo de classificação. Nesse caso, dados que são classificados de maneira errônea por classificadores anteriores têm maior peso no projeto do "classificador presente".
- Tomemos um conjunto com N dados. Associamos a esses dados pesos da forma $w_n, n = 1, \ldots, N$. Agora, se considerarmos que os pesos correspondem à tarefa de projetar a m-ésima máquina, chegamos à notação ${w_n}^{(m)}$, $n = 1, \ldots, N$ e $m = 1, \ldots, M$. A Fig. 3 ilustra, nesse contexto, a estratégia de *boosting*.

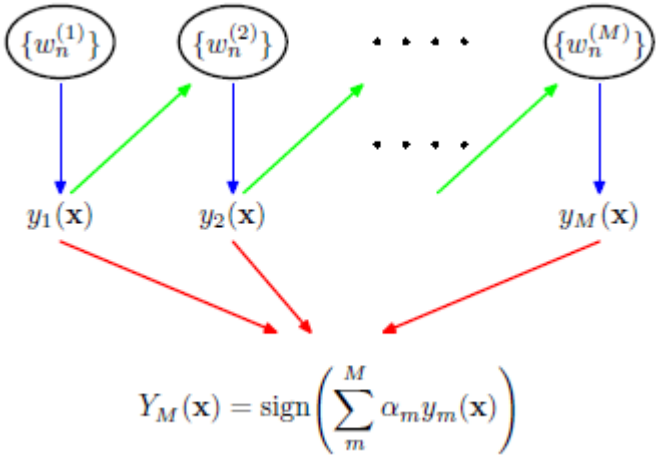

Figura 3. Estratégia de *Boosting* (de [Bishop, 2006]).

- A emblemática metodologia conhecida como *Adaboost* parte de uma escolha "igualitária" para os pesos, ou seja, $w_n^{(1)} = 1/N$, para $n = 1, \ldots, N$. A partir daí, o valor dos pesos na iteração $m + 1$ depende do desempenho de classificação da máquina associada à iteração m: padrões classificados erradamente tendem ter maior peso e padrões corretamente classificados tendem a ter menor peso [Bishop, 2006].

2.1.1. Exemplos

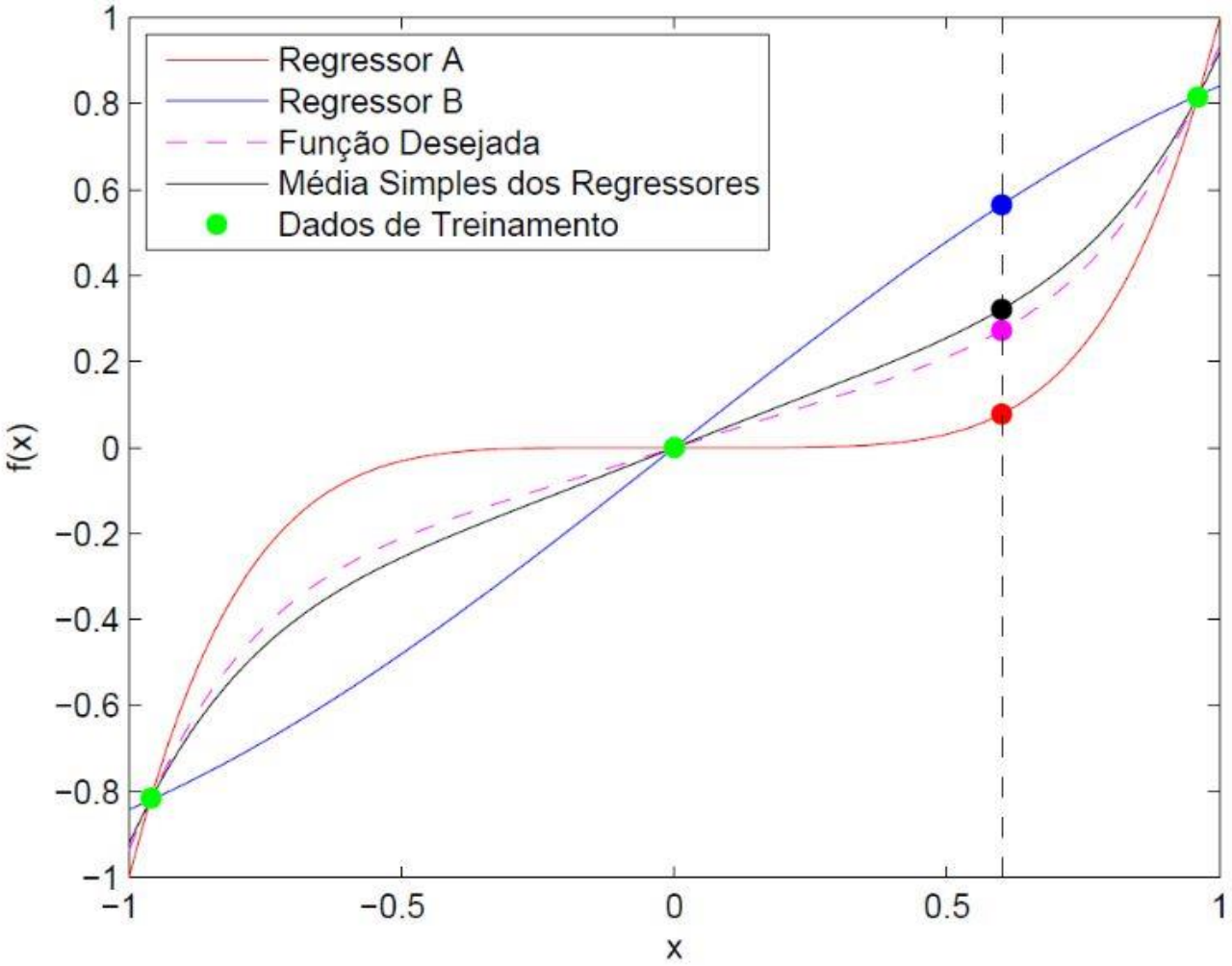

Figura. Aproximação de funções através da combinação de modelos por média simples.

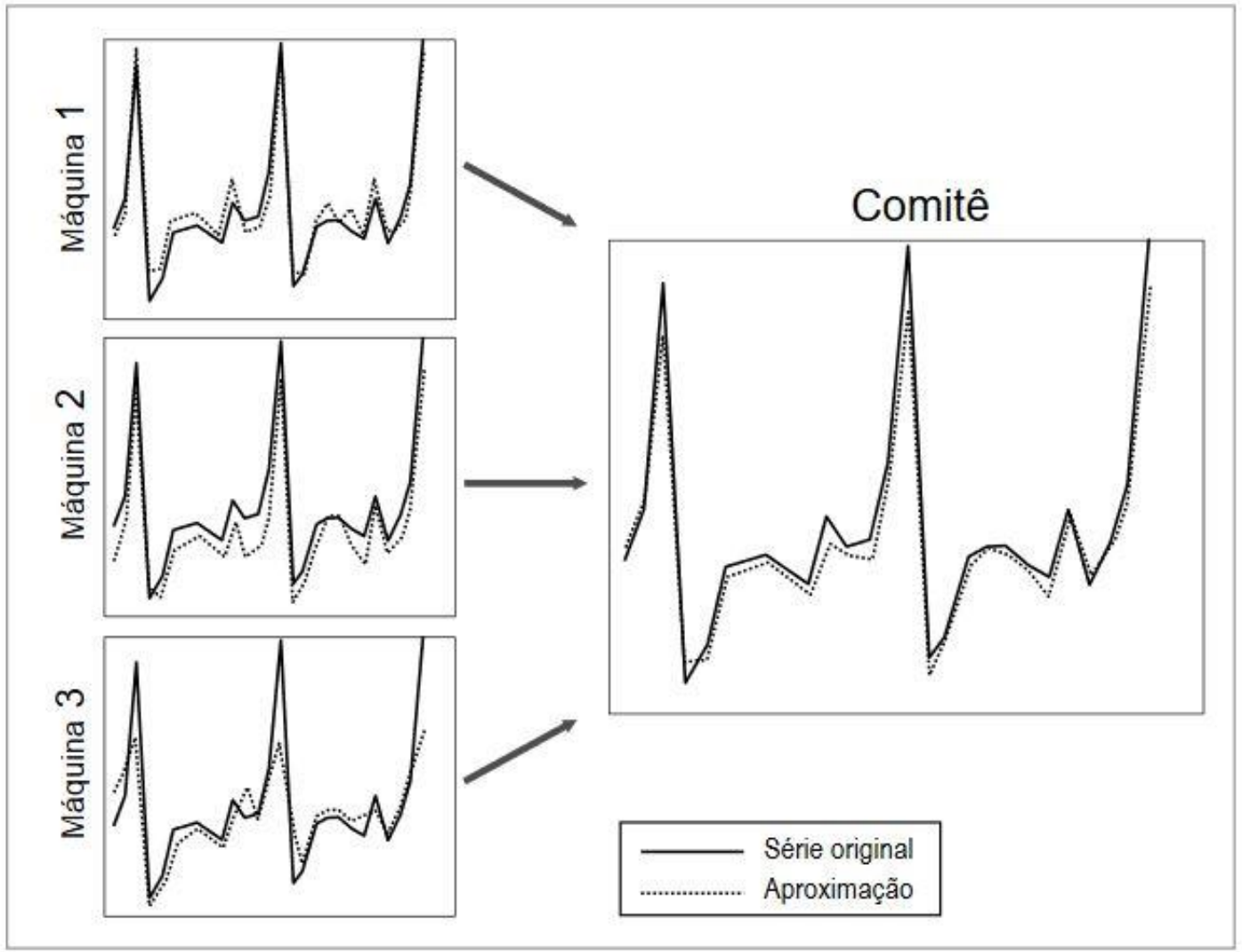

Figura. Combinação por média simples de modelos de previsão de séries temporais.

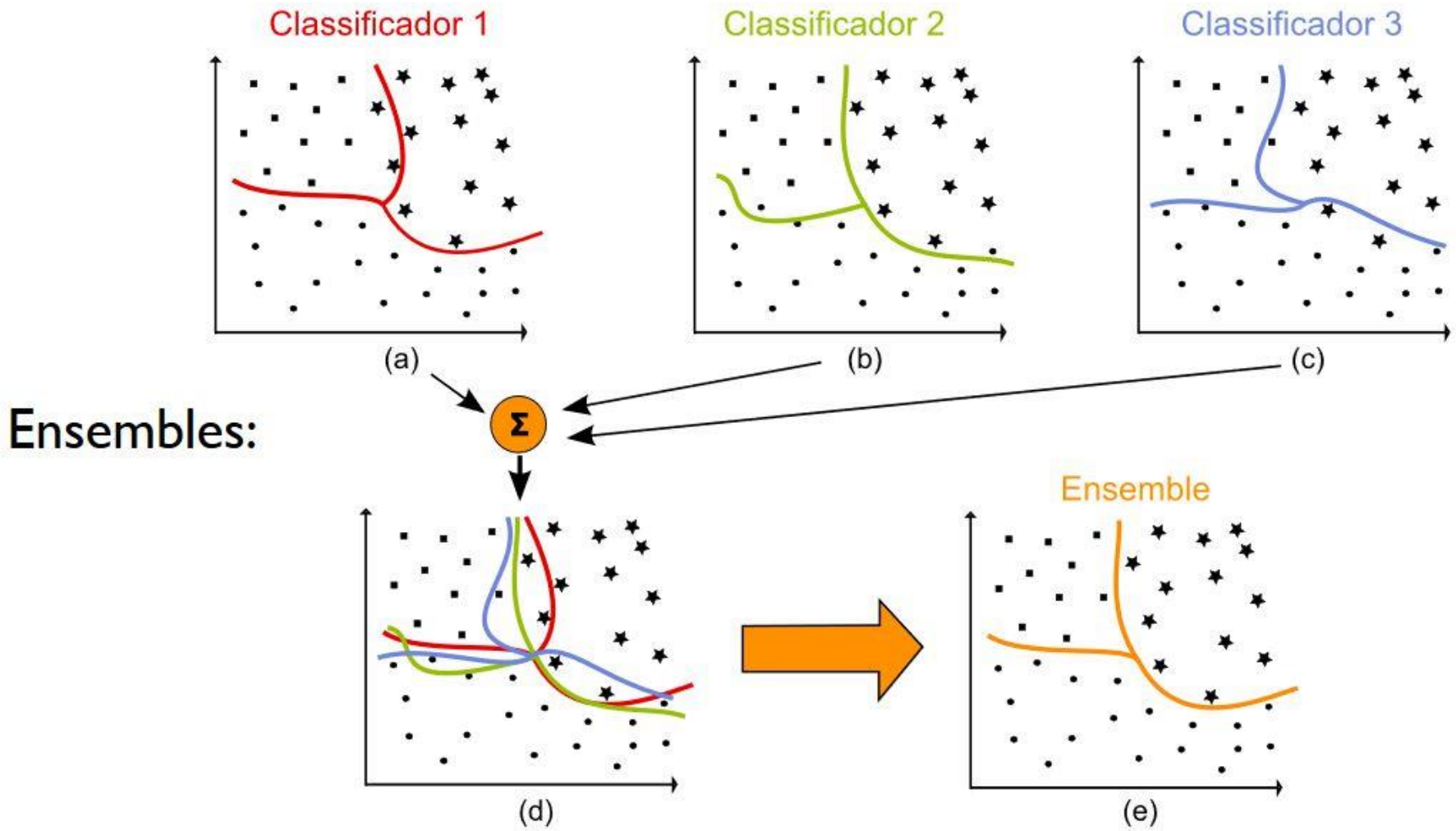

Figura. *Ensemble* de classificadores através de voto majoritário.

2.2. Ensembles: Outros Métodos de Geração de Diversidade

- As abordagens *bagging* e *boosting* geram modelos diversificados quanto aos dados apresentados a cada membro do ensemble. Há, não obstante, outras formas de gerar modelos plurais. Discutiremos a seguir algumas das principais estratégias [Brown et al., 2005].
- O ponto de partida para o treinamento de cada modelo pode ser variado, ou seja, cada modelo pode ter **condições iniciais distintas** no processo iterativo de treinamento. Se forem usadas, por exemplo, redes neurais treinadas com a ajuda do *backpropagation*, diferentes valores aleatórios podem ser utilizados na inicialização. Com isso, em tese, seriam obtidos modelos diversificados. No entanto, estudos empíricos mostram que esse mecanismo não é suficientemente forte em diversos casos de interesse [Brown et al., 2005].

- O **conjunto de arquiteturas usadas** também pode ser variado. Uma possibilidade é usar estruturas idênticas com números distintos de unidades de processamento, e.g., redes MLP com diferentes quantidades de neurônios nas camadas intermediárias. Também podem ser usados diferentes modelos de uma mesma categoria, como redes MLP e RBF. Por fim, podem ser usados modelos de categorias diferentes, como redes neurais e árvores de decisão.
- **Métodos de busca multimodal**, como algoritmos evolutivos com *niching*, também podem ser usados. Nesse caso, realiza-se a otimização da função custo associada ao modelo com uma ferramenta capaz de lidar com múltiplas soluções diversificadas.
- Por fim, mencionamos a possibilidade de que **termos de penalização** sejam introduzidos na função custo utilizada. Esses termos, via de regra, servem para

penalizar configurações em que há elevada correlação entre as estimativas geradas pelos membros do comitê.

2.3. Mistura de Especialistas

- No caso de um comitê dinâmico, como na abordagem conhecida como **mistura de especialistas**, tem-se uma filosofia distinta: o problema é "resolvido por partes" pelos componentes. Cada componente não mais resolve o problema como um todo, mas o **conjunto de componentes** é responsável por isso [Coelho et al., 2016].
- Dessa forma, faz-se necessário projetar os componentes e também o estágio de controle conhecido como *rede gating*. Essa rede indica, a cada entrada, o papel que os componentes do comitê devem desempenhar [Tresp, 2001].

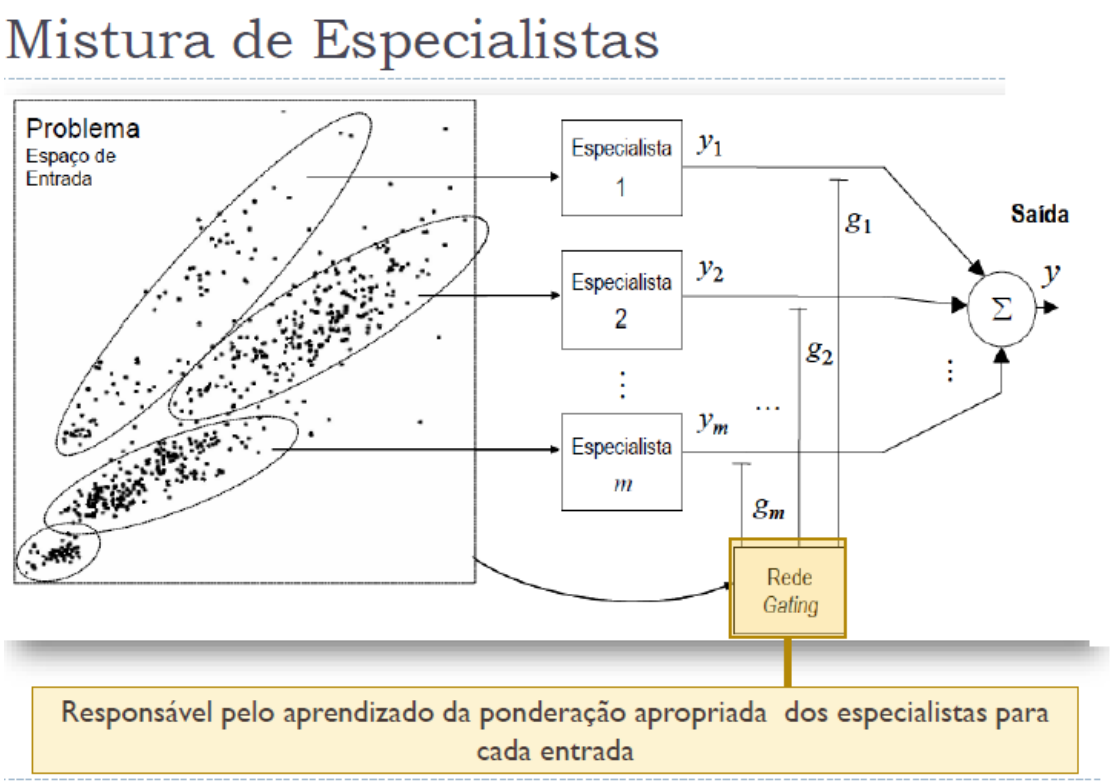

Figura 4. Esquema de Mistura de Especialistas (de [Coelho et al., 2016]).

- A rede *gating* pode ser implementada por uma arquitetura neural (linear ou não-linear) com uma camada de saída *softmax*, que retornaria os pesos $g_m, m = 1, \dots, M$, a serem tratados como parâmetros de uma combinação linear das saídas dos especialistas $(y_1, y_2, \dots, y_M)$. Nesse caso, ter-se-ia algo como:

$$y_C = \sum_{m=1}^{M} g_m y_m$$

- O treinamento, como indicado, deve abranger o ajuste dos parâmetros dos especialistas e da rede *gating* [Coelho et al., 2016].

3. Referências bibliográficas

BISHOP, C. M., Pattern Recognition and Machine Learning, Springer, 2006.

ROWN, G., WYATT, J., HARRIS, R., YAO, X., "Diversity Creation Methods: a Survey and Categorisation", Information Fusion, Vol. 6, No. 1, pp. 5 – 20, 2005.

COELHO, G. P., VON ZUBEN, F. J., ATTUX, R., "Comitês de Máquinas: Ensembles e Misturas de Especialistas", Notas de Aula do Curso IA353 / 2016, Disponíveis em www.dca.fee.unicamp.br/~vonzuben/courses/ia353_1s16.html , 2016.

MARKOWITZ, H., "Portfolio Selection", The Journal of Finance, Vol. 7, No. 1, pp. 77 – 91, 1952.

PUMA-VILLANUEVA, W. J. "Comitês de Máquinas em Predição de Séries Temporais". Dissertação de Mestrado, Faculdade de Engenharia Elétrica e de Computação, UNICAMP, 2006.

TRESP, V., "Committee Machines", *in* Hu, Y. H., Hwang, J.-N. (eds.), *Handbook for Neural Network Signal Processing*, CRC Press, 2001.