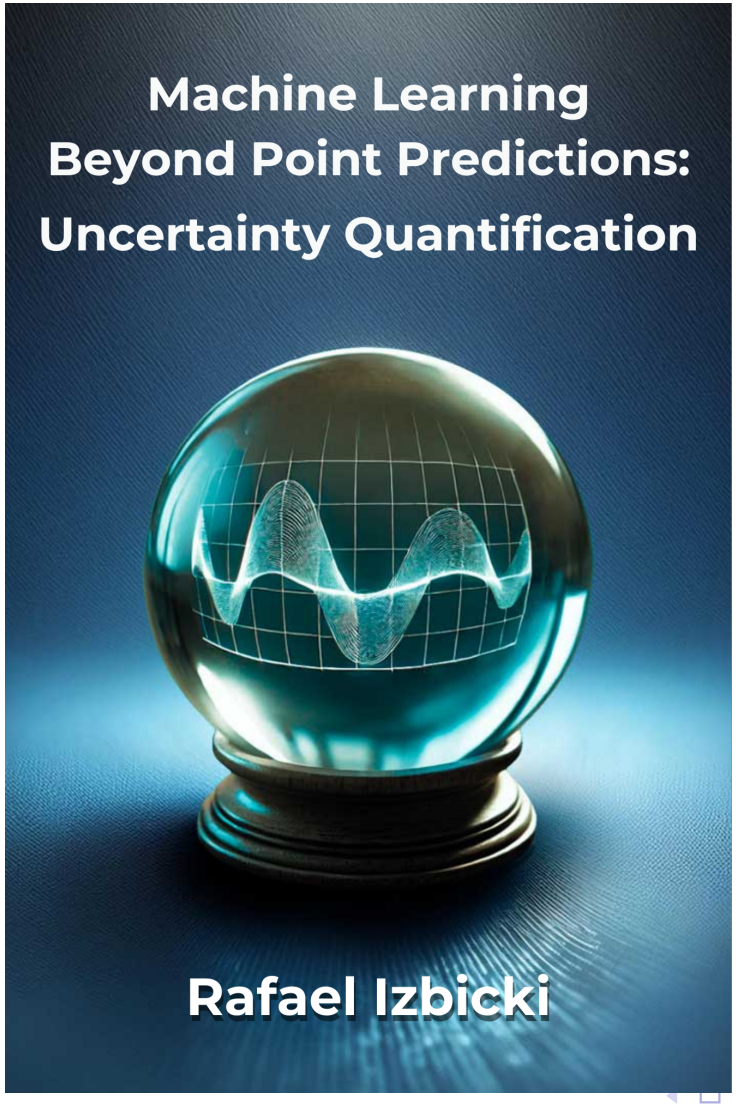
Introdução a Incerteza em Machine Learning

Marcos O Prates

20 de Fevereiro de 2025

Principal Referência

Izbicki, R. Machine Learning Beyond Point Predictions: Uncertainty Quantification. Disponível em: `https://rafaelizbicki.com/uq4mlpt/` (Izbicki, 2025)



Sobre a incerteza

- A incerteza associada à previsão de um novo rótulo Y dado características x pode ser dividida em duas categorias principais:
  - Incerteza Aleatória: Este tipo de incerteza surge quando o mesmo vetor de características x corresponde a diferentes rótulos ou medidas possíveis de y.
  - Incerteza epistêmica: essa incerteza é devida à falta de conhecimento sobre o verdadeiro processo de geração de dados (ou seja, a verdadeira distribuição de $Y|x$).

- Em muitos problemas de previsão, é importante distinguir entre esses dois tipos de incerteza, pois eles exigem tratamento diferente.
  - A incerteza epistêmica pode ser mitigada reunindo conjuntos de dados maiores ou melhorando o modelo.
  - A incerteza aleatória só pode ser reduzida medindo covariáveis adicionais.

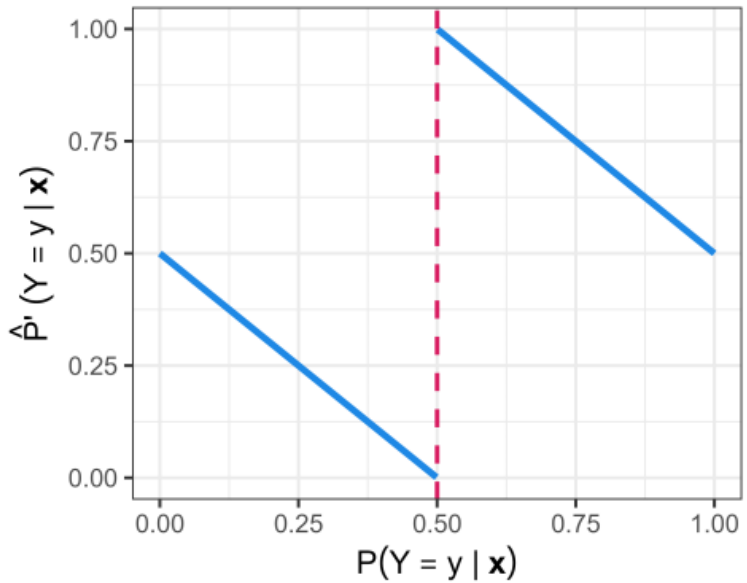
Densidade Condicional

Usando Densidade Condicional

- Em vez de focar em uma única previsão de ponto $g(x)$ para Y, geralmente é mais útil entender a incerteza em torno de Y.
- Vamos focar na estimação da distribuição condicional $f(y|x)$, tanto para respostas discreta como contínuas.

Vamos entender porque uma boa predição pontual não é necessariamente um bom estimador para incerteza.

- Considere dois modelos probabilísticos. O primeiro, $\hat{P}(Y = y|x) = P(Y = y|x)$,
- O segundo, $\hat{P}'(Y = y|x)$, distorce as probabilidades verdadeiras, ou seja, aproxima de 0,5 quando estão perto de 0 ou 1, e os afasta de 0,5 quando estão perto disso.



- Ambos modelos $\hat{P}(Y = y|x)$ e $\hat{P}'(Y = y|x)$ geram a mesma classificação, pois

$$\hat{P}(Y = y|x) \geq 1/2 \text{ if and only if } \hat{P}'(Y = y|x) \geq 1/2.$$

- Note que $\hat{P}'$ distorce as probabilidades, deslocando-as para $0,5$ quando a certeza é alta e para longe de $0,5$ quando a incerteza é maior, falhando assim em quantificar a incerteza adequadamente.

Classificadores Probabilísticos

- O uso de modelos paramétricos são alternativas para estimação da densidade, e.x., regressão linear ou MLG. Nessa abordagem é necessário somente estimar os parâmetros do modelo.
- Porém, uma escolha não realista para a distribuição condicional dos dados podem tornar os modelo paramétricos mau estimadores para a densidade. Nesse sentido, estimados não paramétricos podem ser tornar atrativos.

FlexCode

- Estimativa de densidade condicional não paramétrica flexível via regressão (FlexCode, Izbicki and B. Lee, 2017) é uma alternativa. Disponível em `https://github.com/rizbicki/FlexCoDE`.
- O método introduz uma abordagem totalmente não paramétrica para estimativa de densidade condicional ao reformular o problema como uma série ortogonal, onde a regressão é utilizada para estimar coeficientes de expansão.
- Esta estratégia permite estimativa eficiente de densidades condicionais em altas dimensões, alavancando os sucessos da regressão de alta dimensão.

- Em resumo o método consite em:
  1. Especifique uma base ortonormal $(\phi_i)i \in \mathbb{N}$ in $\mathbb{R}$
  2. For fixed $x \in \mathbb{R}^d$ and $f(\cdot|x)$. Let

$$f(y|x) = \sum_{i \in \mathbb{N}} \beta_i(x)\phi_i(y).$$

  3. Isto implica que, para um i fixo, podemos estimar $\beta_i(x)$ regredindo $\phi_i(y)$ em x usando a seguinte amostra transformada $(X_1, \phi_i(Y_1)), \ldots, (X_n, \phi_i(Y_n))$
  4. Logo,

$$\hat{f}(y|x) \sum_{i=1}^{I} \hat{\beta}_i(x)\phi_i(y).$$

  onde $\hat{\beta}_i(x) = E(\phi_i(Y)|x)$.

Mistura de Modelos

- A motivação por trás deste métodos vem do fato que mistura finitas de normais podem aproximar qualquer densidade (Titterington et al., 1985), desde que o número de componentes seja grande o suficiente.
- Uma mistura de normal é dada por

$$f(y|x) = \sum_{i=1}^{m} \alpha_i(x)\phi(y|\theta_i(x))$$

onde $\alpha_i(x)$'s são pesos não negativos tal que $\sum_{i=1}^{m} \alpha_i(x) = 1$, ϕ é a densidade da normal e $\theta_i(x) = (\mu_i(x), \sigma_i^2(x))$ (Tatiana et al., 2009).

Estimadores de Densidade por Kernel

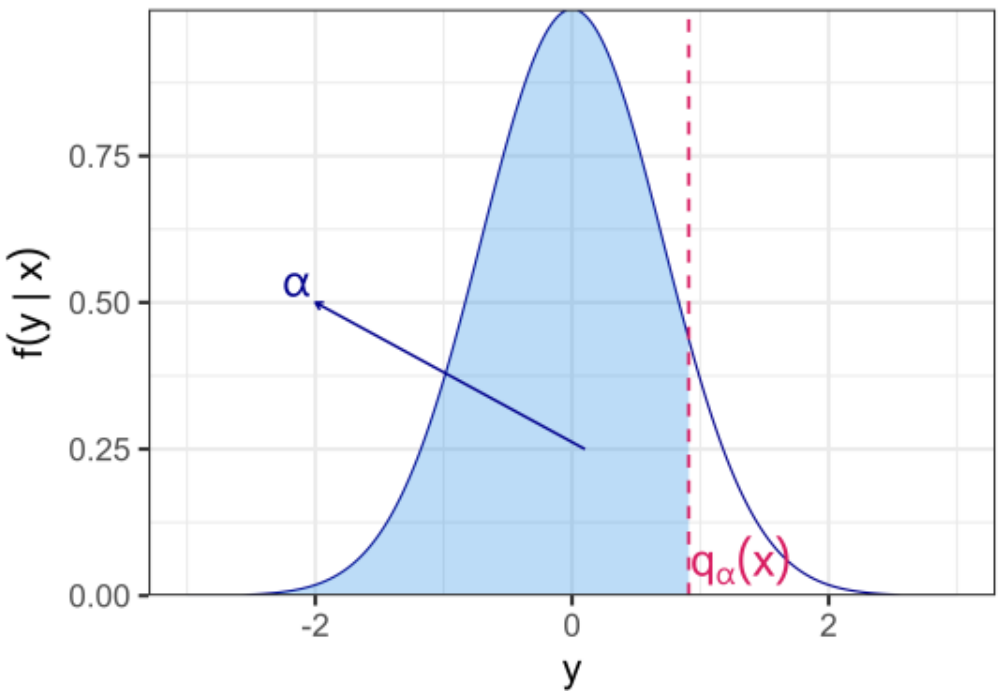
- Para dados de baixa dimensão, estimadores de densidade por Kernel são uma opção (Rosenblatt, 1969)
- Um estimador de Kernel estima $f(y|x) = \frac{f(y,x)}{f(x)}$, usualmente implementada como

$$\hat{f}(y|x) = \frac{\sum_{i=1}^{n} K_{h_x}(\| x - X_i \|) K_{h_y}(\| y - Y_i \|)}{\sum_{i=1}^{n} K_{h_x}(\| x - X_i \|)}$$

onde $K_h(t) = h^{-d} K(t/h)$ é um kernel com largura de banda h em dimensão d (Hayfield and Racine, 2008).

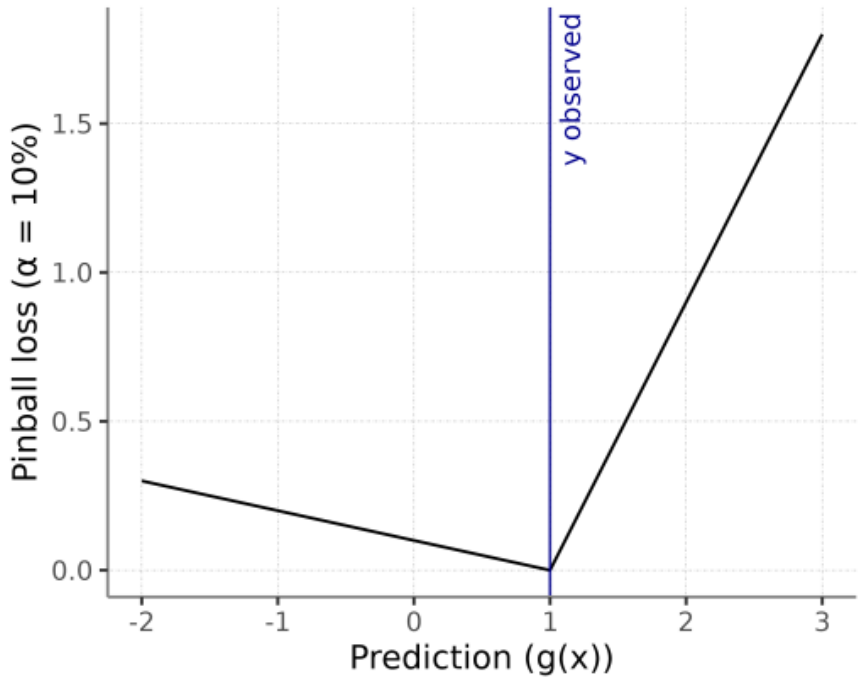
Regressão Quantílica

- A regressão quantílica visa estimar uma propriedade da distribuição de $Y|x$: seus quantis condicionais.
- Formalmente, o quantil $\alpha - \textit{condicional}$ de Y em x, $q_\alpha(x)$, é a função tal que $P(Y \leq q_\alpha(x)|X = x) = \alpha$.



- Para avaliar a precisão de uma estimativa de $q_\alpha(x)$, usaremos a função de perda chamada de perda de pinball, dada por

$$L_\alpha(g, x, y) = (g(x) - y)(\mathbb{I}(y \leq g(x)) - \alpha).$$



Métodos para estimar regressão quantílica

- **Método paramétricos**: Uma maneira de estimar $q_\alpha(x)$ é assumir que ele depende linearmente das covariáveis (Koenker and Bassett Jr, 1978):

$$q_\alpha(x) = \beta^\top x$$

- **KNN**: A ideia do método k-vizinhos mais próximos (KNN) para estimar $q_\alpha(x)$ é usar as respostas Y dos k-vizinhos mais próximos para x (Ma et al., 2016). Formalmente temos

$$g(x) = \hat{q}_\alpha(x) = (\{y_i\}_{i \in \mathcal{N}_x})$$

onde $q_\alpha(x) = (\mathcal{S})$ é o quantil-α de $\mathcal{S}$ e $\mathcal{N}_x$ são o k-vizinhos mais próximos de x. (Meinshausen and Ridgeway, 2006)

Métodos para estimar regressão quantílica

- **Florestas aletórias**: Estima quantis a partir de uma estimativa da distribuição cumulativa condicional (Meinshausen and Ridgeway, 2006).
- Esta estimativa é uma versão local da estimativa dada pela função de distribuição cumulativa empírica:

$$\hat{F}(y|x) = \sum_{i=1}^{n} w_i(x)\mathbb{I}(y_i < y)$$

onde $w_i(x)$ é uma medida de similaridade entre x e a i-ésima observação no conjunto de treinamento obtido por meio de uma floresta aleatória.

Métodos para estimar regressão quantílica

- Especificamente, seja R_x^b a folha onde a observação x cai na b-ésima árvore, e defina o peso associado a essa árvore como

$$w_i(x, b) = \frac{\mathbb{I}(x_i \in R_x^b)}{\sum_{j=1}^{n} \mathbb{I}(x_j \in R_x^b)}$$

- Em outras palavras, $w_i(x, b)$ é a proporção de observações no conjunto de treinamento que caem na mesma folha que x. O peso combinado é então dado pela média desses pesos:

$$w_i(x) = \frac{1}{B} \sum_{b=1}^{B} w_i(x, b).$$

- Este método tem várias vantagens sobre o KNN
  - Realiza automaticamente a seleção de variáveis;
  - Leva em conta interações entre inputs;
  - levam em conta não linearidades dos inputs.

Regiões de Incerteza

Métodos para estimar regioes de incerteza

- **Região de mario densidade preditiva (HPD))**: A região HPD inclui ps valores mais prováveis de y e equilibra a precisão com o tamanho da região. Formalmente temos

$$R(x) = \{y \in \mathcal{Y} : f(y|x) > \lambda\}$$

onde $\lambda > 0$ controla a troca entre incluir o y em R e manter R pequeno.

Métodos para estimar regioes de incerteza

- **Região quantílica**: Fixe $\alpha \in (0, 1)$e assuma que a região de previsão é um intervalo, $R(x) = (a(x), b(x))$. A região de predição ótima é o intervalo baseado em quantil:

$$R(x) = (q_{\alpha/2}(x), q_{1-\alpha/2}(x))$$

onde $q_\gamma(x) = F^{-1}(\gamma|x)$ representa o γ-ésimo quantil condicional de Y dado x.

Métodos para estimar regioes de incerteza

- **Região simétrica**: Considere o caso em que $\mathcal{Y} \subset \mathbb{R}$. Fixe $\lambda > 0$ e assuma que a região de predição é um intervalo, $R(x) = (a(x), b(x))$. A região de predição ótima é

$$R(x) = \mathbb{E}(Y|x) \pm \sqrt{\lambda \mathbb{V}(Y|x)}$$

onde $\mathbb{E}(Y|x)$ representa a média condicional de Y dado x, e $sqrt\mathbb{V}(Y|x)$ representa o desvio padrão condicional de Y dado x.

Cada região de previsão tem seus próprios pontos fortes e fracos:

- **HPD**: Essas regiões são particularmente eficazes para capturar os resultados mais prováveis, tornando-as adequadas para distribuições assimétricas ou multimodais. No entanto, elas podem resultar em regiões de formato irregular que são computacionalmente desafiadoras para lidar/relatar.
- **RQ**:Intervalos baseados em quantis são simples de computar e interpretar, exigindo apenas dois números para representar o intervalo. No entanto, essas regiões podem levar a regiões amplas em casos multimodais.
- **RS**:Regiões simétricas são fáceis de calcular e são particularmente eficazes para distribuições simétricas, também precisando de apenas dois números para representação. No entanto, elas podem se tornar grandes em distribuições assimétricas.

Regiões usando Plug-ins

- Uma abordagem direta e simples para construir regiões de predição envolve aproveitar uma estimativa inicial da função de densidade condicional, $\hat{f}$, e inseri-la diretamente na expressão para a região ótima. Por exemplo, HPD, quantílica ou simétrica.
- As regiões de predição de plug-in são projetadas principalmente para capturar a incerteza aleatória inerente aos dados.
- No entanto, uma vez que as quantidades como a densidade condicional $f(y|x)$ ou os quantis $q_{\alpha/2}(x)$ são estimados apenas a partir de dados, as regiões de predição resultantes geralmente não alcançam a cobertura correta.

Regiões Conformes

Regiões Conformes

- O principal objetivo em previsões conformes é usar os dados para obter uma região de predição válida, isto é, uma região de predição cuja cobertura corresponde à sua cobertura nominal, $R(X_{n+1})$, sob muito poucas suposições.
- De fato, tipicamente, a validade depende apenas da suposição i.i.d. (Angelopoulos et al., 2020; Shafer and Vovk, 2008)

- Uma abordagem amplamente empregada para construir tais regiões de predição é o método split (Lei et al., 2018; Vovk, 2012).
- Neste método, os dados são divididos em dois conjuntos: o conjunto de treinamento, D_1, e o conjunto de calibração, D_2. Assumimos que o tamanho do conjunto de calibração é $|D_2| = n$.

- Após os dados terem sido divididos, uma pontuação de não conformidade $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ é treinada usando D_1.
- A função $h(x, y)$ é usada para quantificar a extensão em que um dado valor de rótulo y se alinha com os valores de caracaterística x de uma instância.
- Um alto valor de $h(x, y)$ sugere que o aparecimento do rótulo y em uma instância caracterizada por caracaterísticas x é improvável.

- **[Divisão de regressão]** $h(x, y) = |y - \hat{r}(x)|$, onde $\hat{r}$ é uma estimativa da regressão de Y em x (Lei et al., 2018). Isso mede a diferença absoluta entre o valor observado y e o valor previsto $\hat{r}(x)$, indicando o quanto o resultado real se desvia da previsão da regressão e é essencialmente o resíduo da regressão.
- **[Ponderado]** $h(x, y) = \frac{|y - \hat{r}(x)|}{\hat{\rho}(x)}$, onde $\hat{\rho}(x)$ é qualquer estimador do desvio absoluto médio condicional de $|Y - \hat{r}(X)||X = x$ (Lei et al., 2018). Esta pontuação normaliza a diferença absoluta entre y e o valor previsto $\hat{r}(x)$ pelo desvio esperado em torno da predição, tornando a pontuação de não conformidade relativa à variabilidade aleatória nos dados.

- **[CD-split e variações]** $h(x, y) = -\hat{f}(y|x)$ (Izbicki et al., 2022). Isso mede a não conformidade por meio da densidade condicional estimada de y dado x, o que significa que uma densidade menor indica um resultado menos provável.
- **[HPD-split]** $h(x, y) = -\int_{\{y':\hat{f}(y'|x)\leq\hat{f}(y|x)\}} \hat{f}(y'|x)dy'$ (Izbicki et al., 2022). Esta pontuação quantifica o quão longe y está das regiões de maior densidade, destacando o quão incomum y é em relação a outros resultados possíveis.

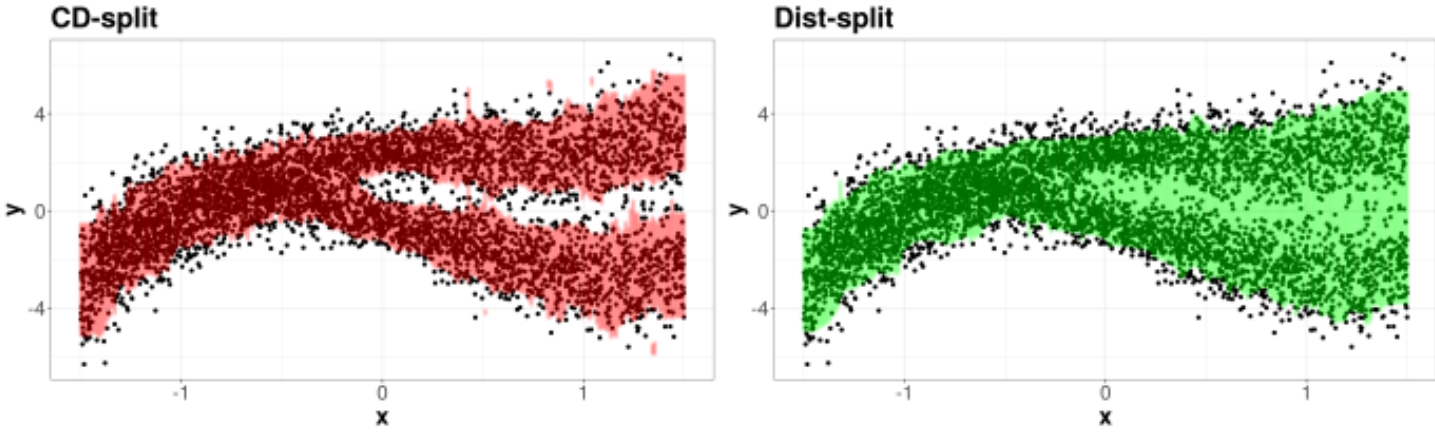
- **[CQR]** $h(x, y) = \max\{\hat{q}_{\alpha_1}(x) - y, y - \hat{q}_{\alpha_2}(x)\}$, onde $\hat{q}_{\alpha_1}$, $\hat{q}_{\alpha_1}$ são estimativas de quantil e $\alpha_1 < \alpha_2$ (Romano et al., 2019). Isso mede o quão longe y está das estimativas de quantil, capturando o quanto y se desvia dos limites de quantil inferior e superior previstos para x.
- **[CDF-split]** $h(x, y) = |\hat{F}(y|x) - 1/2|$, onde $\hat{F}$ é uma estimativa de CDF (Chernozhukov et al., 2021). Isso mede a distância entre o valor de CDF de y e $1/2$, indicando o quão centralizado y está dentro da distribuição de resultados possíveis, com valores próximos a $1/2$ sendo mais típicos.

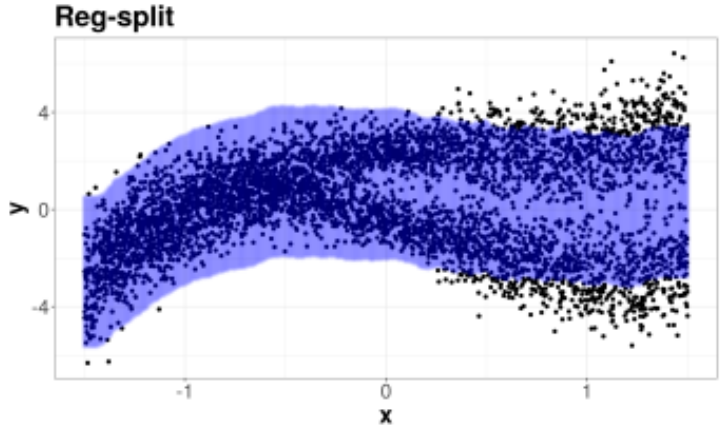
- A região conforme tem a forma

$$R(x_{n+1}) = \{y : h(x_{n+1}, y) < t\},$$

  o que significa que consiste em todos os y que são altamente conformes a x.
- A pontuação de não conformidade h determina a forma e as propriedades da região de predição.
- Por exemplo, a regressão-divisão sempre leva a intervalos homocedásticos centrados em torno de $\hat{r}(x)$,

$$R(x_{n+1}) = (\hat{r}(x_{n+1}) - t, \hat{r}(x_{n+1}) + t)$$


CD-split
Dist-split
4
0
-4
y
-1
0
1
x


Reg-split
4
0
-4
y
-1
0
1
x

Estimadores Bayesianos

Bayesian Additive Regression Trees (BART)

- Na sua forma mais simples, o BART (Chipman et al., 2010) assume que

$$Y = h(x) + \epsilon = \sum_{b=1} g_b(x) + \epsilon$$

  onde $\epsilon \sim N(0, \sigma^2)$.
- Especificamente, $g_b(x)$ assume a forma de uma árvore de regressão.

- Seja T_b a topologia da árvore de regressão associada a g_b, e seja $\mu_b = \{\mu_{b,1}, \ldots, \mu_{b,|T_b|}\}$ as saídas associadas com cada uma das $|T_b|$ folhas de T_b. Logo

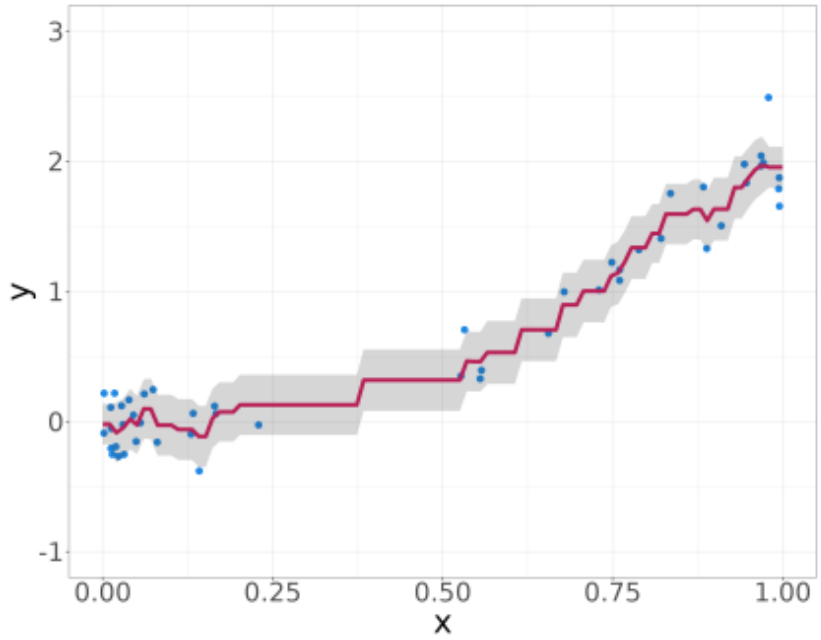
$$g_b(x) = \mu_{b,i}$$

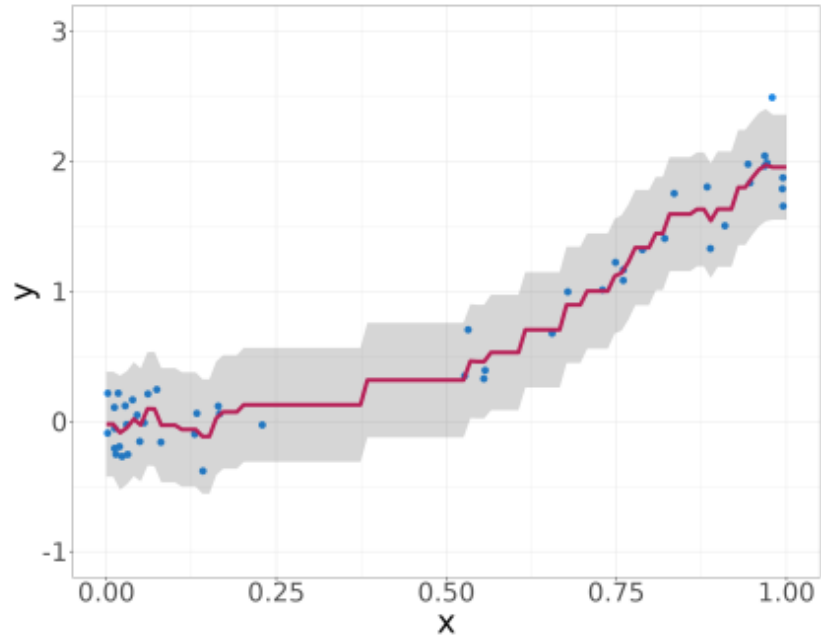
onde $\mu_{b,i}$ é a saída de T_b associada a x.

- O modelo BART assume, portanto, que a função de regressão é uma soma das saídas de cada árvore de regressão

$$\mathbb{E}(Y|x) = \sum_{b=1}^{B} g_b(x) = \sum_{b=1}^{B} \mu_{b,i}.$$

Os parâmetros do modelo são, portanto $(T_1, \mu_1, \ldots, T_B, \mu_B, \sigma)$ (Sparapani et al., 2021).


3
2
1
0
-1
y
0.00
0.25
0.50
0.75
1.00
x


3
2
1
0
-1
y
0.00
0.25
0.50
0.75
1.00
x

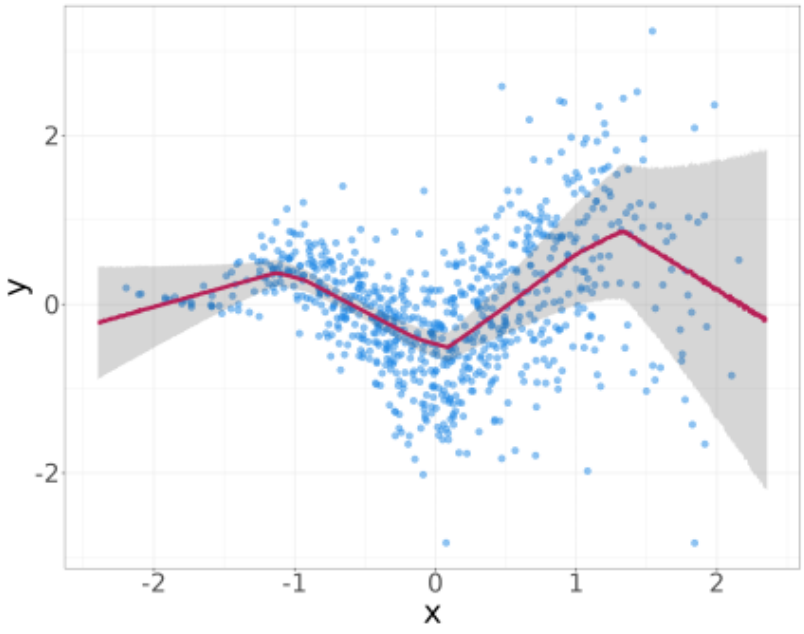
Monte Carlo Dropout

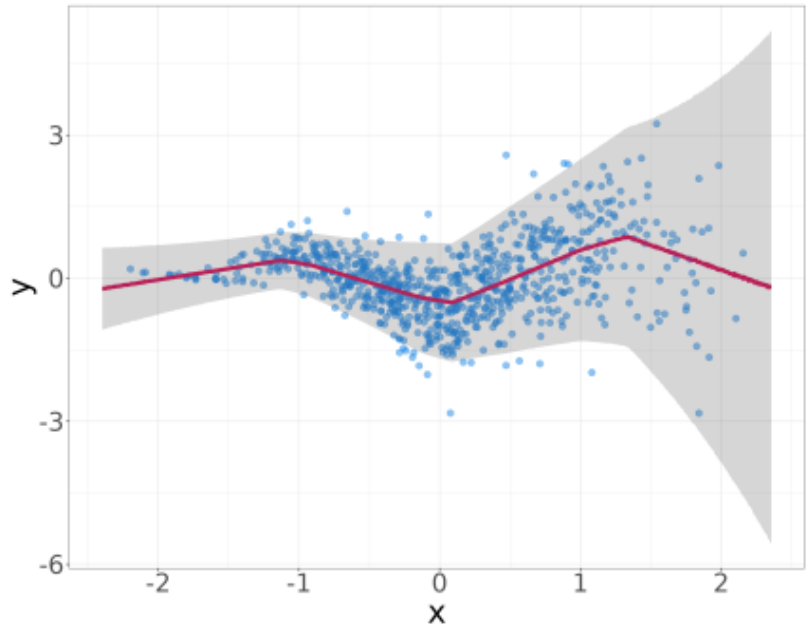
- Monte Carlo Dropout é uma técnica originalmente projetada para evitar overfitting de uma rede neural (Srivastava et al., 2014), mas isso também pode ser usado para medir a incerteza em torno de previsões pontuais em tais redes.
- Em palavras, o dropout define aleatoriamente a função de ativação de cada nó como zero com probabilidade p.
- Se usado durante o tempo de previsão, o dropout induz uma distribuição sobre as possíveis saídas de uma rede ajustada.
- This randomness can be used to quantify the uncertainty around the predictions.

- Se um modelo para a incerteza aleatória, $Y|x, \theta$, estiver disponível, podemos aproximar a distribuição preditiva Bayesiana via

$$f(y, |x, D) \approx \int f(y, |x, \theta) q(\theta) d\theta$$

- Podemos efetivamente amostrar a partir da distribuição preditiva por:
  1. Amostragem θ de $q(\theta)$. Isso pode ser feito usando o dropout de Monte Carlo na rede.
  2. Amostragem de $Y|x, \theta$ usando o modelo para a incerteza aleatória.


2
0
-2
y
-2
-1
0
1
2
x


3
0
-3
-6
y
-2
-1
0
1
2
x

Referências I

Angelopoulos, A., S. Bates, J. Malik, and M. I. Jordan (2020). Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*.

Chernozhukov, V., K. Wüthrich, and Y. Zhu (2021). Distributional conformal prediction. *Proceedings of the National Academy of Sciences 118*(48), e2107794118.

Chipman, H. A., E. I. George, and R. E. McCulloch (2010). Bart: Bayesian additive regression trees.

Hayfield, T. and J. S. Racine (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software 27*(5), 1–32.

Izbicki, R. (2025). *Machine Learning Beyond Point Predictions: Uncertainty Quantification* (1st ed.).

Izbicki, R. and A. B. Lee (2017). Converting high-dimensional regression to high-dimensional conditional density estimation.

Referências II

Izbicki, R., G. Shimizu, and R. B. Stern (2022). Cd-split and hpd-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research 23*(87), 1–32.

Koenker, R. and G. Bassett Jr (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Lei, J., M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association 113*(523), 1094–1111.

Ma, X., X. He, and X. Shi (2016). A variant of k nearest neighbor quantile regression. *Journal of Applied Statistics 43*(3), 526–537.

Meinshausen, N. and G. Ridgeway (2006). Quantile regression forests. *Journal of machine learning research 7*(6).

Referências III

Romano, Y., E. Patterson, and E. Candes (2019). Conformalized quantile regression. *Advances in neural information processing systems 32*.

Rosenblatt, M. (1969). Conditional probability density and regression estimators. *Multivariate analysis II 25*, 31.

Shafer, G. and V. Vovk (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research 9*(3).

Sparapani, R., C. Spanbauer, and R. McCulloch (2021). Nonparametric machine learning and efficient computation with bayesian additive regression trees: The bart r package. *Journal of Statistical Software 97*, 1–66.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research 15*(1), 1929–1958.

Referências IV

Tatiana, B., C. Didier, R. David, and Y. Derek (2009). mixtools: an r package for analyzing finite mixture models. *Journal of Statistical Software 32*(6), 1–29.

Titterington, D. M., A. F. Smith, and U. E. Makov (1985). Statistical analysis of finite mixture distributions. *(No Title)*.

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR.