Covariate and Label Shifts

Semana 08 - Aula 01
Flavio Figueiredo

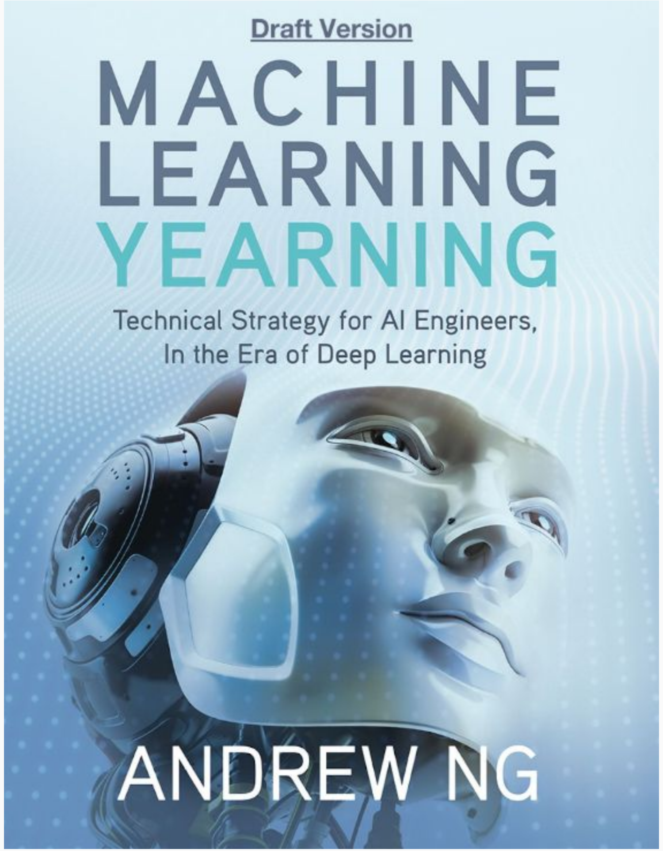
Até Agora

- Geralmente estamos assumindo train once, use always
- Porém o mundo real é bem mais complicado
- Como fazer uso de algoritmos de aprendizado máquina em produção?!
- Problemas:
  - Escalabilidade (não abordado, mas posso conversar)
  - Engenharia de Software (acredito ser menos crítico no escopo de vocês)
    - Feedback loops
    - Relacionado com hoje
  - Shifts em Dados
    - Aula de Hoje

Motivação

Test and Validation Sets

- Chapter 6:
  Your dev and test sets should come from the same distribution
- Essa é uma premissa clássica do aprendizado de máquina
- Mas o mundo real não é tão controlado assim



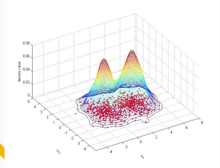
O Pipeline Comum

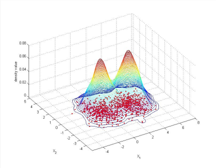
Dados

Treino

Val (Dev)

Teste

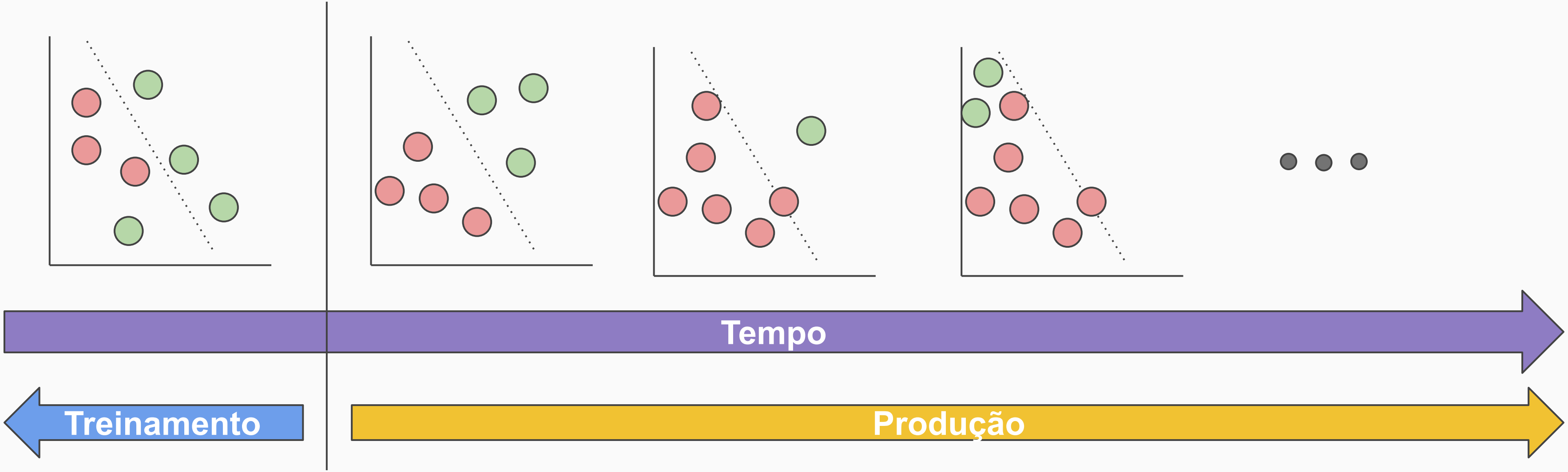




Embora não falamos de treino, para o modelo funcionar a mesma premissa vale. O Andrew está mais focado em boas avaliações

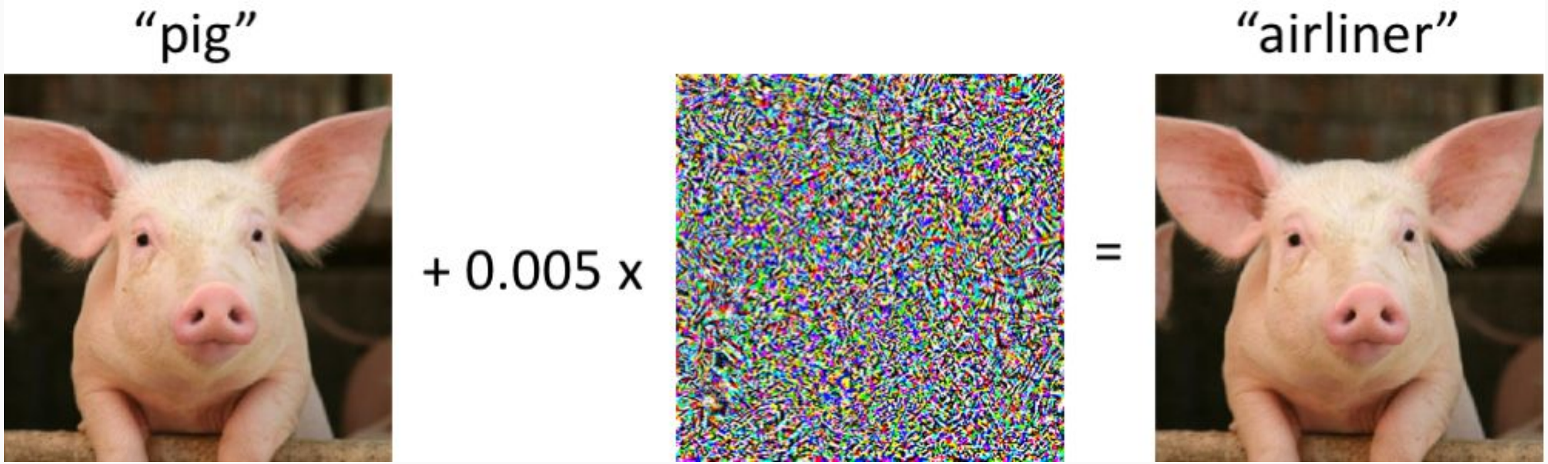
Até Agora

- Geralmente estamos assumindo train once, use always
- Porém o mundo real é bem mais complicado
- Abaixo aprendemos alguma região de decisão (via perceptron ou SVM)
  - Treino, Validação e Teste como esperado
  - Com o passar do tempo. . .



Mudanças nos Dados

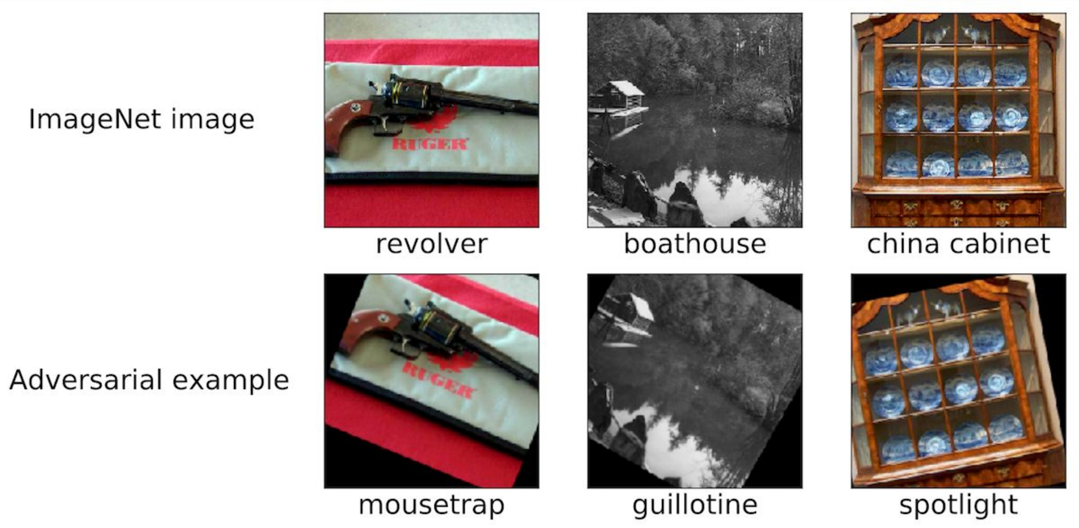
- Podemos pensar casos onde é raro
  - Detecção de dígitos (MNIST)
  - Detecção de faces
- Mas tais problemas não representam a complexidade do mundo real
- Mesmo em casos simples temos problemas



https://gradientscience.org/intro_adversarial/

Mudanças nos Dados

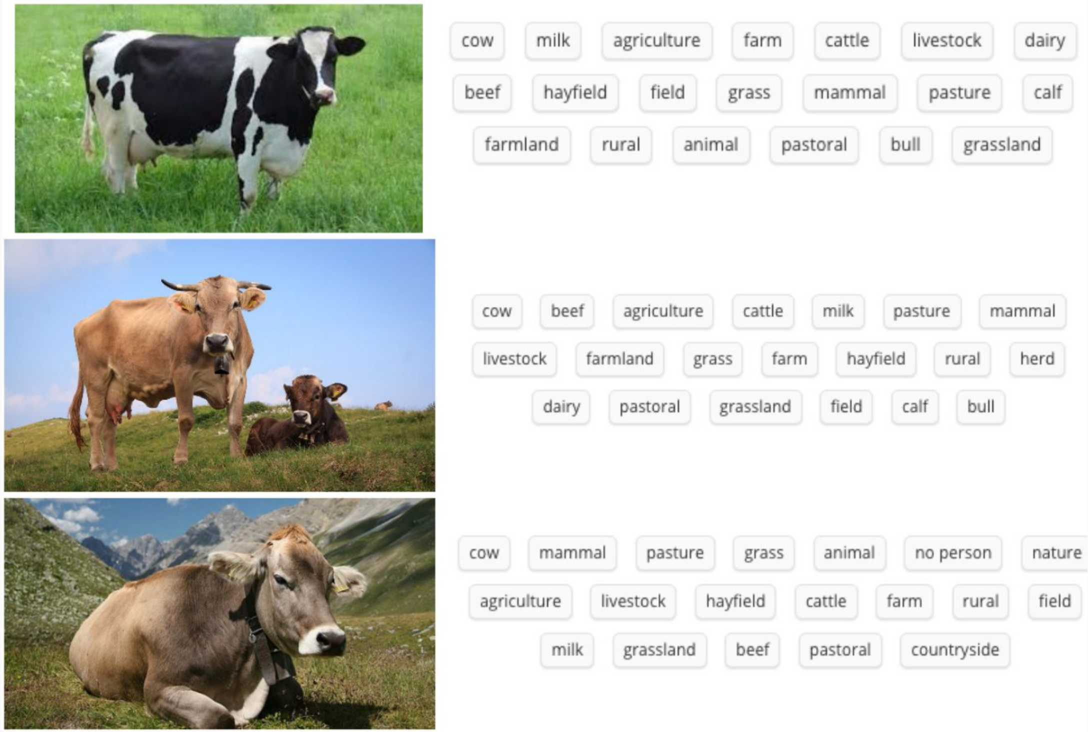
- Para entender um pouco melhor o problema podemos ir para o mundo de Adversarial Machine Learning
- Perguntas sobre a robustez de algoritmos na presença de ruídos



https://gradientscience.org/intro_adversarial/

Sistema Clarifai

- O mundo de causalidade também tem bastante exemplos
- Exemplos de Pietro Perona (http://www.vision.caltech.edu/Perona.html)



Sistema Clarifai

- Agora, vamos mudar o local das vacas.

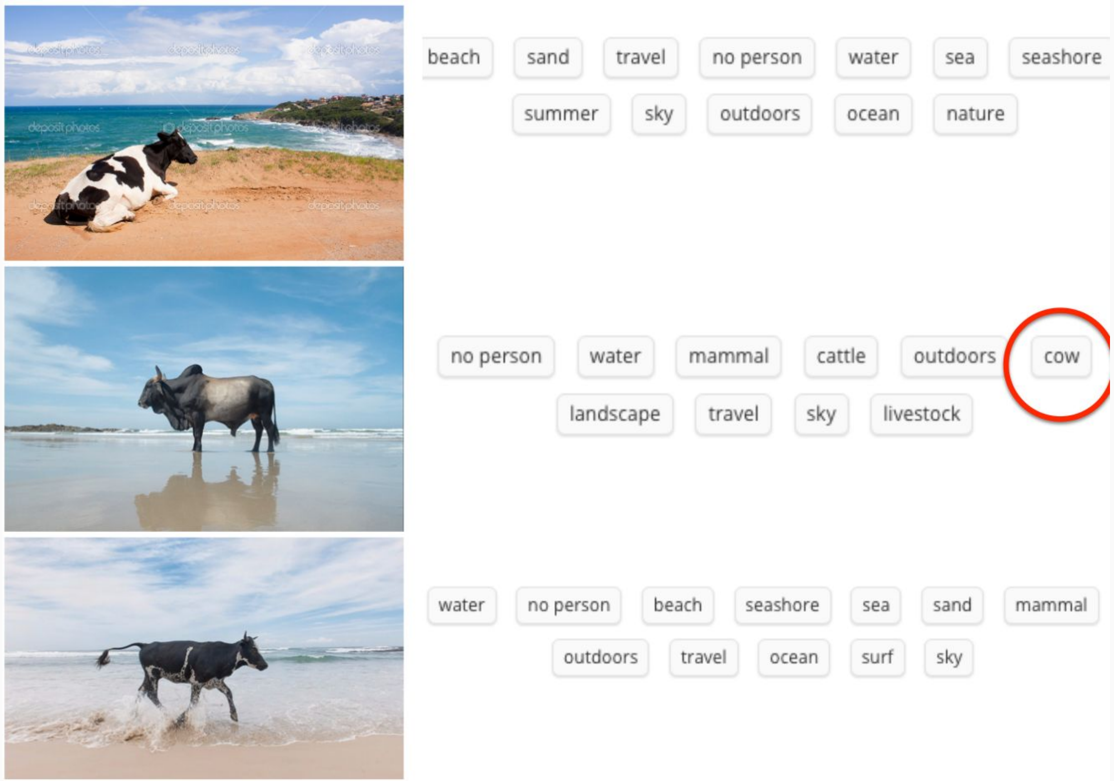


Imagem de uma Vaquinha Suíça (Primeiro ou Segundo do Google)

Meus Testes

clarifai

PRODUCTS ENTERPRISE DEVELOPERS COMPANY DEMO PRICING LOG IN

MORE MODELS

General

VIEW DOCS

LANGUAGE

English (en)

PREDICTED CONCEPT	PROBABILITY
pasture	0.997
livestock	0.995
cow	0.994
grass	0.993
farmland	0.993
agriculture	0.993
cattle	0.993
mammal	0.991
milk	0.991
rural	0.988
beef cattle	0.986
pastoral	0.982

G1: Vacas Curtem Praia em Montenegro

Meus Testes

clarifai

PRODUCTS ENTERPRISE DEVELOPERS COMPANY DEMO PRICING LOG IN

MORE MODELS

General VIEW DOCS

LANGUAGE

English (en)

PREDICTED CONCEPT	PROBABILITY
beach	0.991
water	0.982
sea	0.979
sand	0.978
ocean	0.969
seashore	0.957
travel	0.954
fun	0.954
no person	0.945
mammal	0.945
action	0.944
recreation	0.929

Acontece Botucatu: Vaca é Flagrada Andando Calmamente na Cidade

clarifai

PRODUCTS ENTERPRISE DEVELOPERS COMPANY DEMO PRICING LOG IN

Meus Testes

COLOR

MORE MODELS

General

VIEW DOCS

LANGUAGE

English (en)

PREDICTED CONCEPT	PROBABILITY
cow	0.996
cattle	0.990
livestock	0.986
farm	0.981
milk	0.979
bull	0.975
mammal	0.972
animal	0.967
agriculture	0.948
landscape	0.944
horn	0.931
people	0.909

Porém, observe as outras classes

Meus Testes

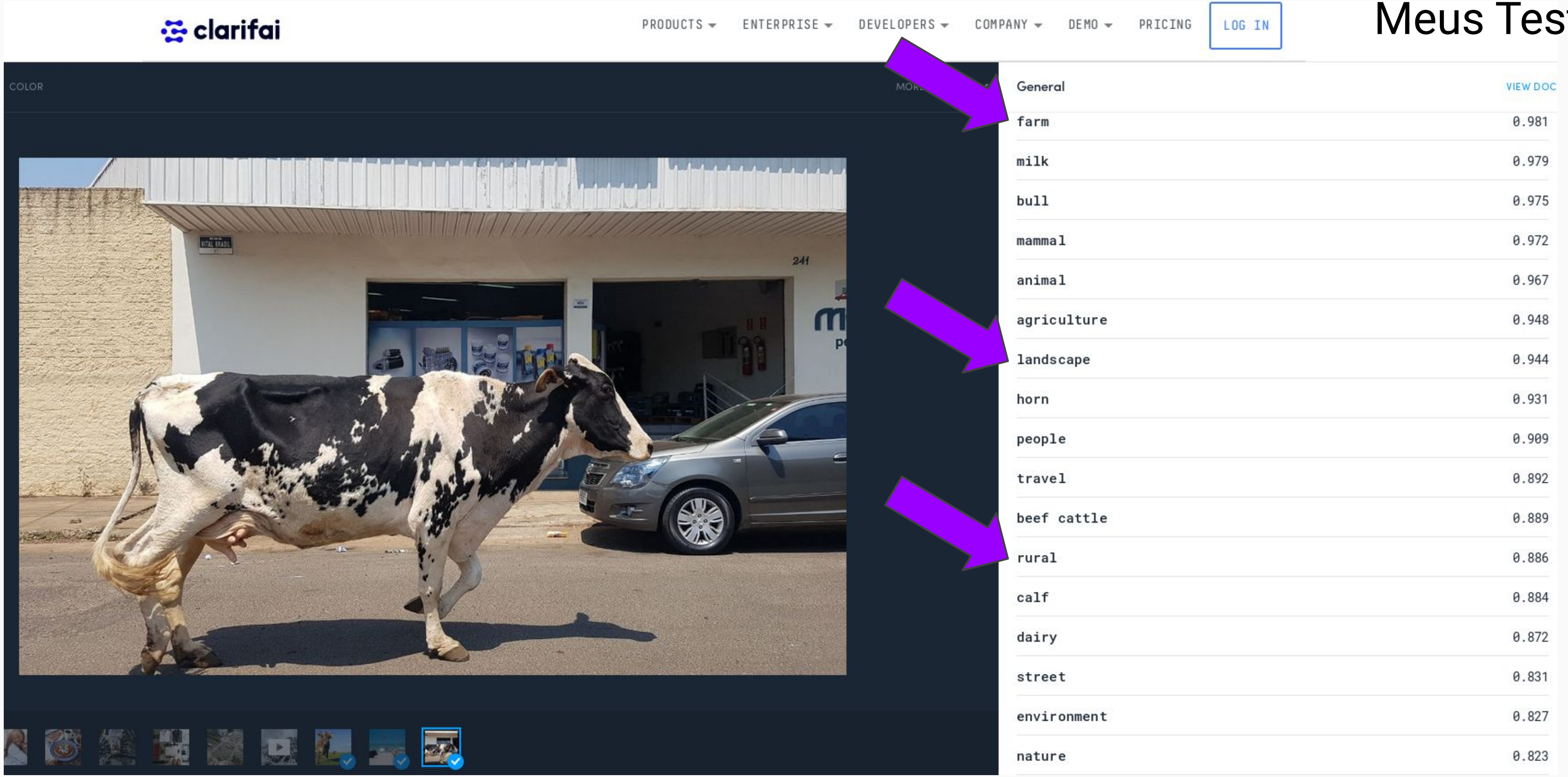
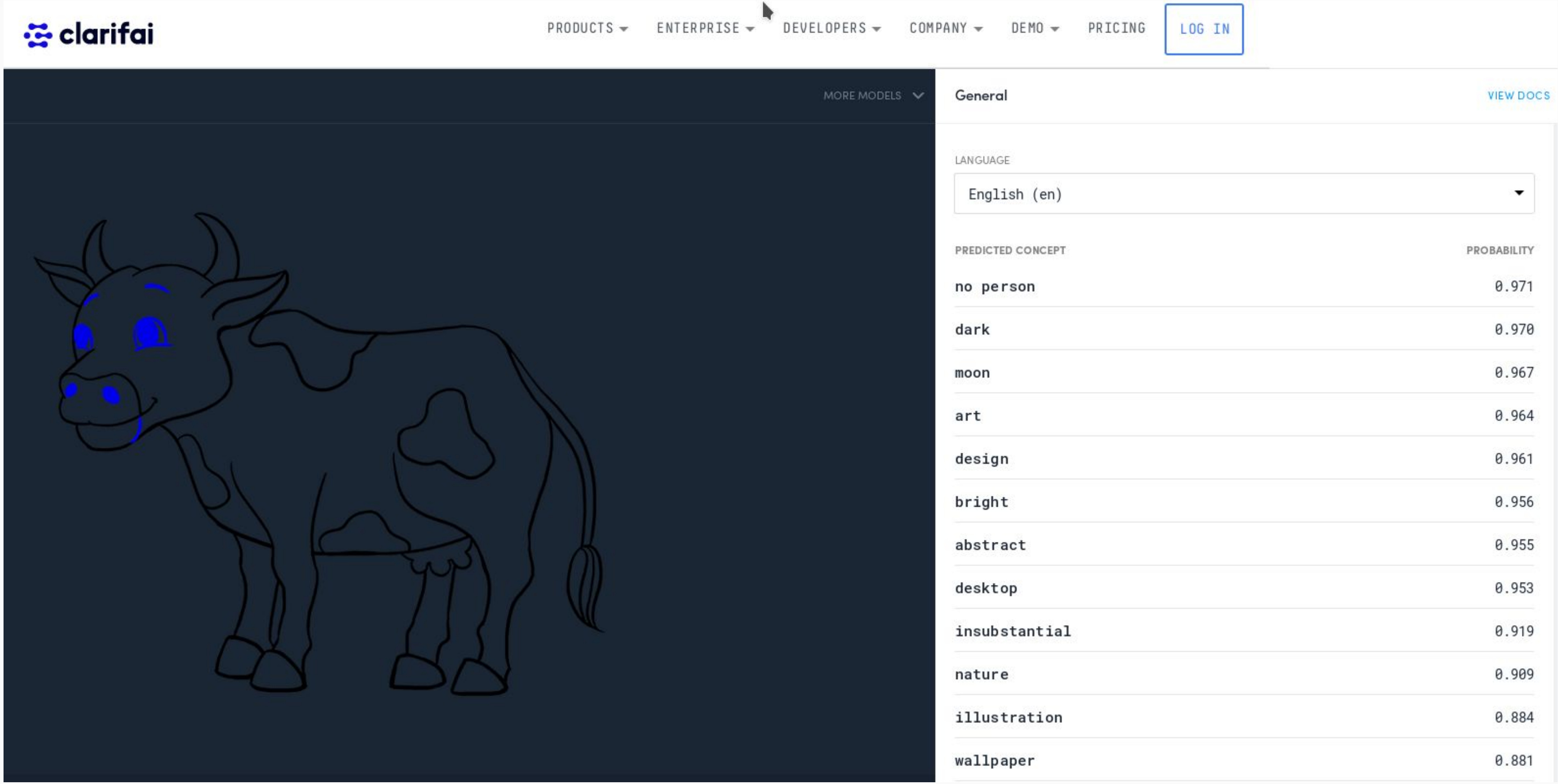
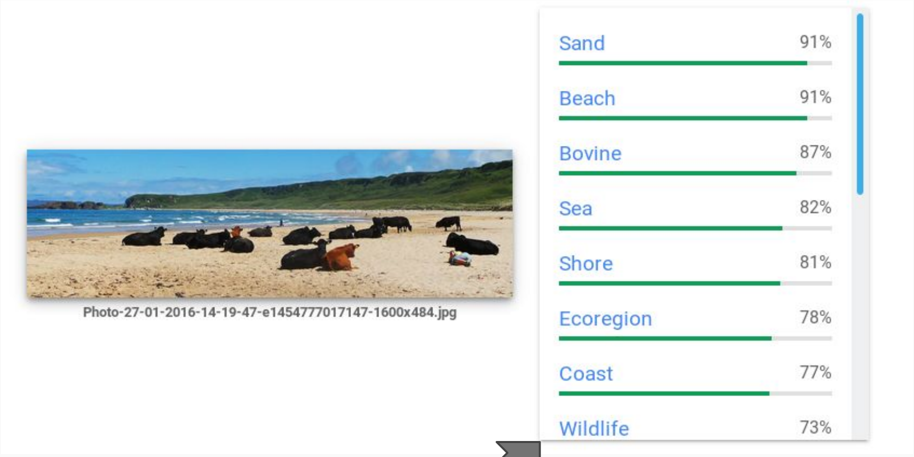


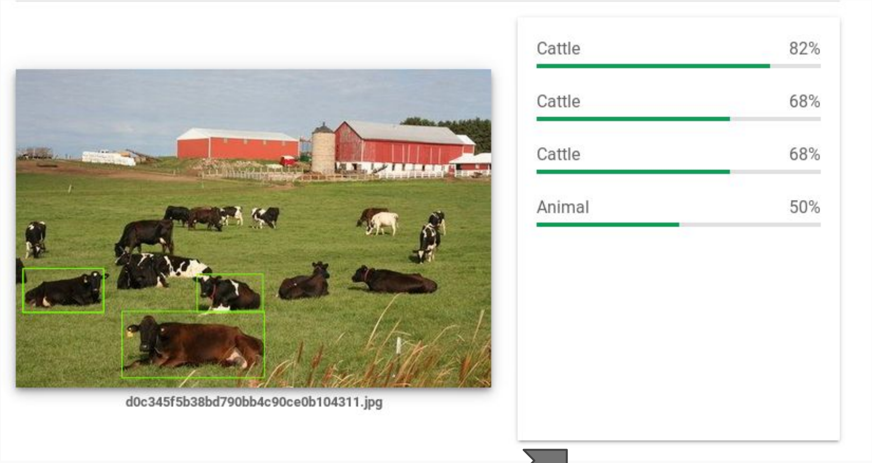
Imagem com fundo transparente



Google Vision API

- Testando outros modelos e empresas famosas
- Nenhuma vaquinha foi detectada
- Agora tentando uma imagem um pouco similar em uma fazenda





Zero objetos

Três objetos

Causalidade e Anti-Causalidade

Causal e Anti-causal

- Embora não seja nosso foco principal. A área de causalidade nos ajuda a entender shifts em dados
- Em particular vamos pensar em dois casos
    - Os atributos mudam
    - Os labels mudam
- Existe um terceiro
    - Ambos mudam, ou o processo gerador de dados como um todo muda
    - Aqui, ainda, nos resta chorar :-(

Causal e Anti-causal

- Sendo X nosso conjunto de atributos.
  - Pixels da imagem
  - Caracteres do texto
  - Tabelinha csv
- Sendo Y nossa resposta
  - Problemas de classificação ou regressão

Causal e Anti-causal

- Sendo X nosso conjunto de atributos.
  - Pixels da imagem
  - Caracteres do texto
  - Tabelinha csv
- Sendo Y nossa resposta
  - Problemas de classificação ou regressão
- Podemos representar uma instância como um vetor **x**. A resposta como y
- Então, estamos trabalhando em um mundo de uma distribuição multivariada
- Também temos que assumir alguma dependência

$$p(\mathbf{x}, y) \neq p(\mathbf{x})p(y)$$

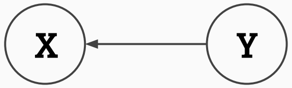
Causal e Anti-causal

- Uma coisa interessante é que a associação entre X e Y pode surgir de diferentes formas. Nosso algoritmo de aprendizado tenta extrair um modelo para tal associação.

- Aqui temos o caso X causa Y
  - Uma bactéria causa pneumonia
  - Clicar no interruptor causa uma luz acender
  - Por si só, pessoas na sala não causam um acendimento das luzes

Causal e Anti-causal

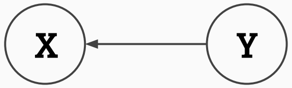
- Uma coisa interessante é que a associação entre X e Y pode surgir de diferentes formas. Nosso algoritmo de aprendizado tenta extrair um modelo para tal associação.



- Aqui temos o caso Y causa X
- Embora parece menos interessante, pense nos dados do curso
  - Imagens
  - Séries temporais
  - Um raio-x causa um câncer de pneumonia? **Não!**

Causal e Anti-causal

- Uma coisa interessante é que a associação entre X e Y pode surgir de diferentes formas. Nosso algoritmo de aprendizado tenta extrair um modelo para tal associação.

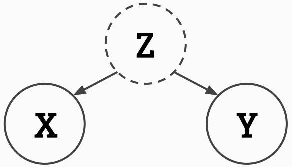


- Aqui temos o caso Y causa X
- Embora parece menos interessante, pense nos dados do curso
  - Imagens
  - Séries temporais
  - Um raio-x causa um câncer de pneumonia? **Não!**

https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

Causal e Anti-causal

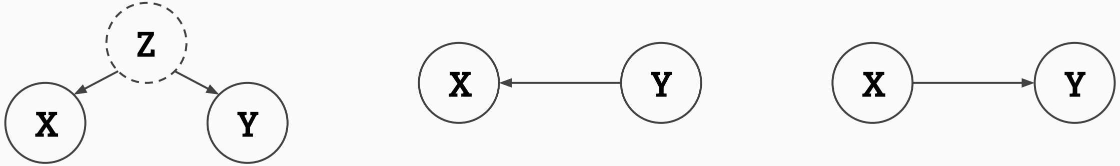
- Uma coisa interessante é que a associação entre X e Y pode surgir de diferentes formas. Nosso algoritmo de aprendizado tenta extrair um modelo para tal associação.



- Por fim temos o caso de um fator latente causando X e Y
- Por exemplo, um aumento de inflação vai ao mesmo tempo
  - Aumentar o preço do ônibus
  - Aumentar o preço de alimentos
  - Os dois são relacionados de uma forma não causal

Reinbach Common Cause Principle

Os três casos



levam para as nossas observações

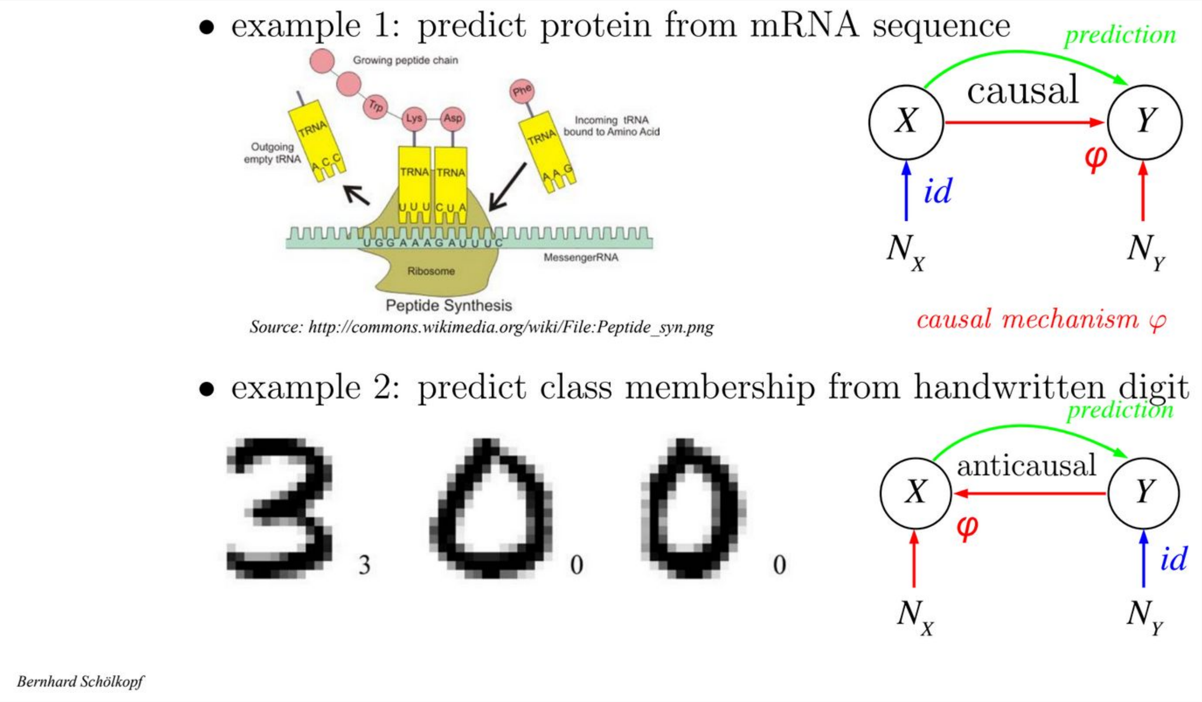
$$p(\mathbf{x}, y)$$

nossos dados de treino são amostras da distribuição conjunta acima

Causalidade vs Previsão

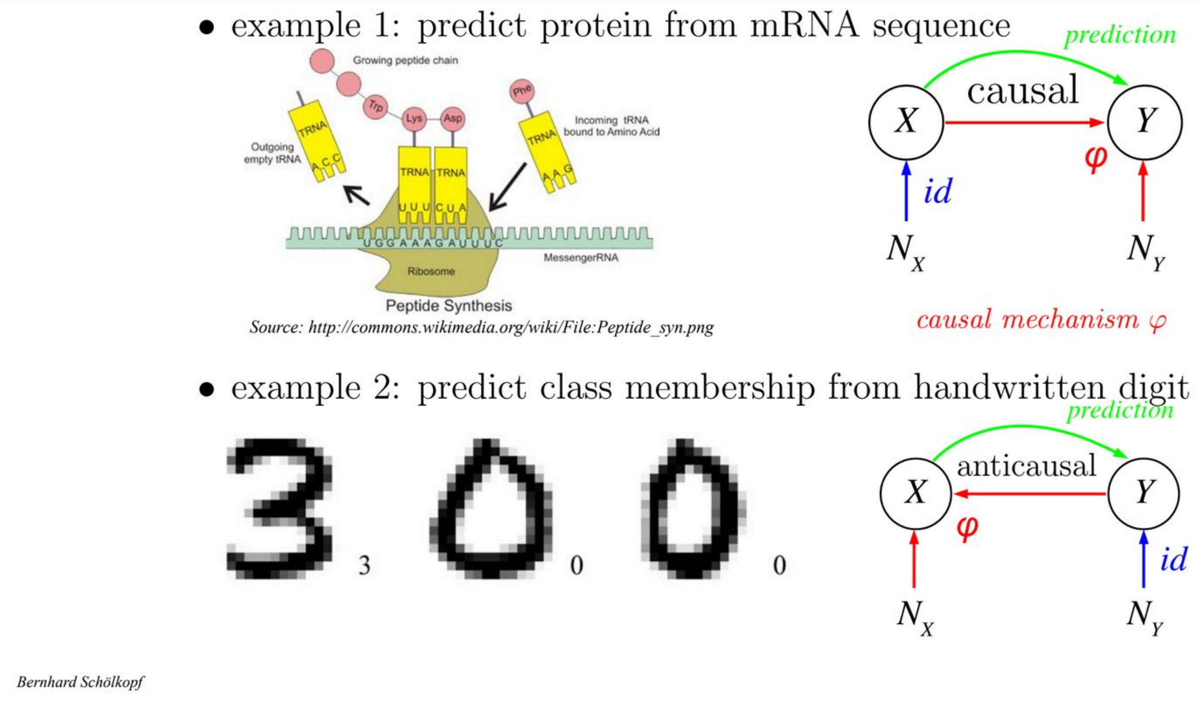
Dois exemplos do Elements of Causal Inference (FREE!). Berhnhard Schölkopf

https://mitpress.mit.edu/books/elements-causal-inference



Causalidade vs Previsão

Note que ainda podemos realizar o aprendizado em casos anti-causais.
O seu modelo vai assumir o contrário, mas ok! Se funciona, ótimo!



On Causal and Anti-Causal Learning (Mais Comum!)

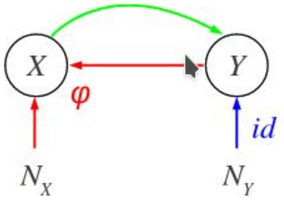


Figure 3. Predicting cause Y from effect X.

$P_{N_Y}(Y - \phi(X))$, where the index N_Y indicates the variable this distribution refers to.

3. Predicting Cause from Effect

We now turn to the opposite direction, where we consider the effect as input and we try to predict the value of the cause variable that led to it. This situation, that we refer to as *anticausal prediction*, may seem unnatural, but it is actually ubiquitous in machine learning. Consider, for instance,

https://icml.cc/2012/papers/625.pdf

Problemas do Curso

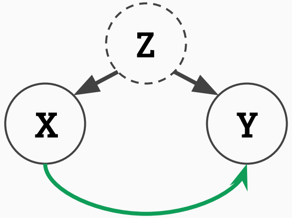
Pergunta: Dos problemas levantados durante o curso, quais seriam um aprendizado **causal** e quais seriam **anti-causal?**

Causalidade vs Inferência

Pequenas Mudanças na Notação!

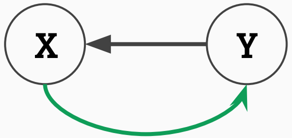
- Observe como causalidade e inferência são correlatos
- Porém separando a terminologia fica mais claro qual é qual
- Existe um processo causal na parte superior de cada figura
- Podemos gerar hipóteses para prever y de x em qualquer um dos casos

$$\mathbf{x}_i = f_{1,\mathbf{\Phi}}(z_i)$$
$$y_i = f_{2,\mathbf{\Phi}}(z_i)$$



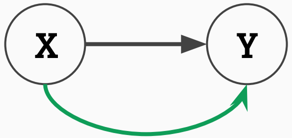
$$\hat{y}_i = h_{\mathbf{\Theta}}(\mathbf{x}_i)$$

$$\mathbf{x}_i = f_{\mathbf{\Phi}}(y_i)$$



$$\hat{y}_i = h_{\mathbf{\Theta}}(\mathbf{x}_i)$$

$$y_i = f_{\mathbf{\Phi}}(\mathbf{x}_i)$$



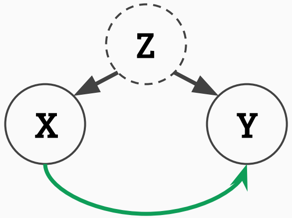
$$\hat{y}_i = h_{\mathbf{\Theta}}(\mathbf{x}_i)$$

Causalidade vs Inferência

Pequenas Mudanças na Notação!

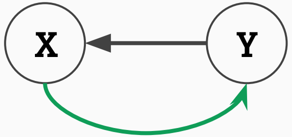
- O mundo do aprendizado profundo é o inferior
  - Geramos hipóteses de modelos que capturam os dados
- Como não sabemos nada do processo causal
  - Pode ser que a hipótese seja inválida no futuro
  - Como proceder?

$$\mathbf{x}_i = f_{1,\mathbf{\Phi}}(z_i)$$
$$y_i = f_{2,\mathbf{\Phi}}(z_i)$$



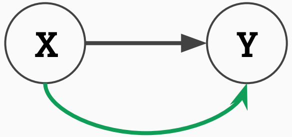
$$\hat{y}_i = h_{\mathbf{\Theta}}(\mathbf{x}_i)$$

$$\mathbf{x}_i = f_{\mathbf{\Phi}}(y_i)$$



$$\hat{y}_i = h_{\mathbf{\Theta}}(\mathbf{x}_i)$$

$$y_i = f_{\mathbf{\Phi}}(\mathbf{x}_i)$$

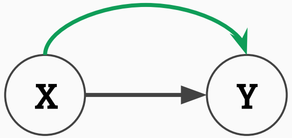


$$\hat{y}_i = h_{\mathbf{\Theta}}(\mathbf{x}_i)$$

Causalidade e Problemas de Aprendizado

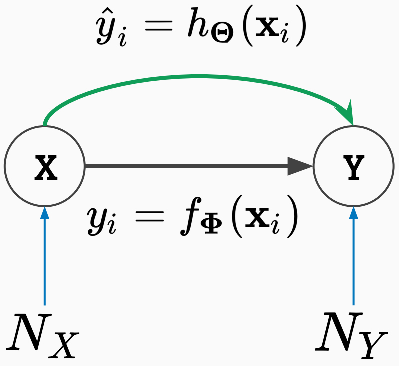
- O exemplo ao lado é um caso causal

$$\hat{y}_i = h_{\mathbf{\Theta}}(\mathbf{x}_i)$$



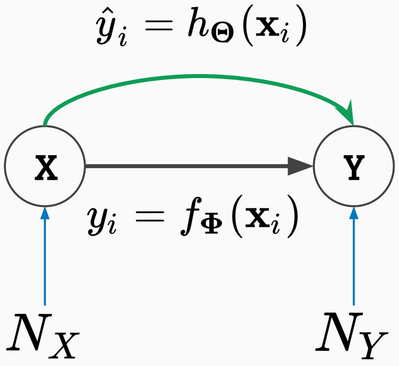
Causalidade e Problemas de Aprendizado

- O exemplo ao lado é um caso causal
- Podemos ter duas fontes de ruído para X e Y
- Nosso objetivo é estimar f através de h
- Assumimos que os dois ruídos são independentes

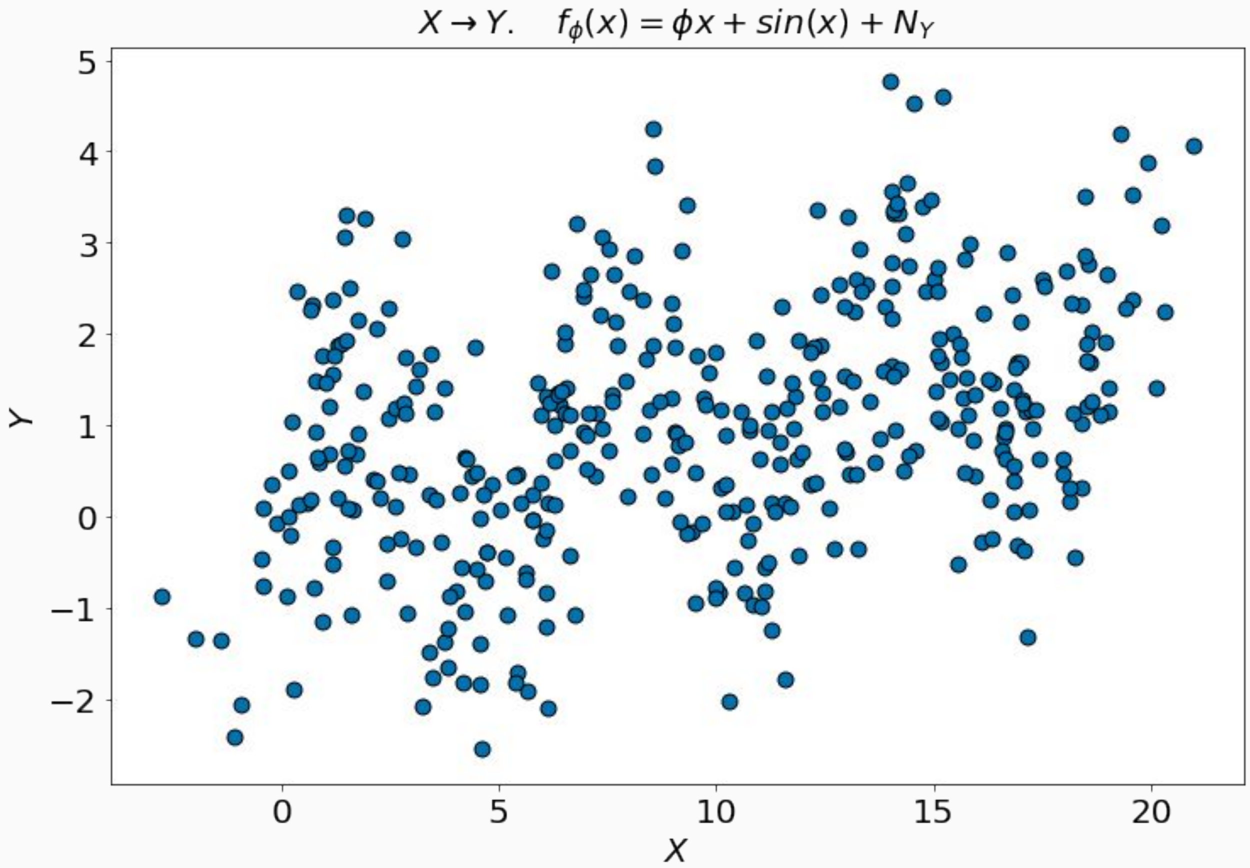


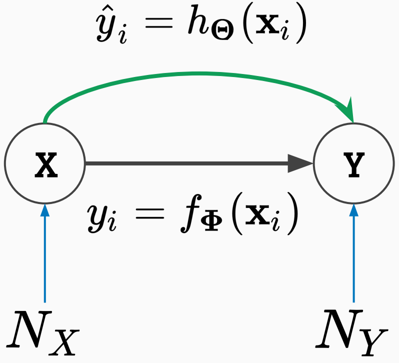
Causalidade e Problemas de Aprendizado

- O exemplo ao lado é um caso causal
- Podemos ter duas fontes de ruído para X e Y
- Nosso objetivo é estimar f através de h
- Assumimos que os dois ruídos são independentes
- Vamos criar um modelo real

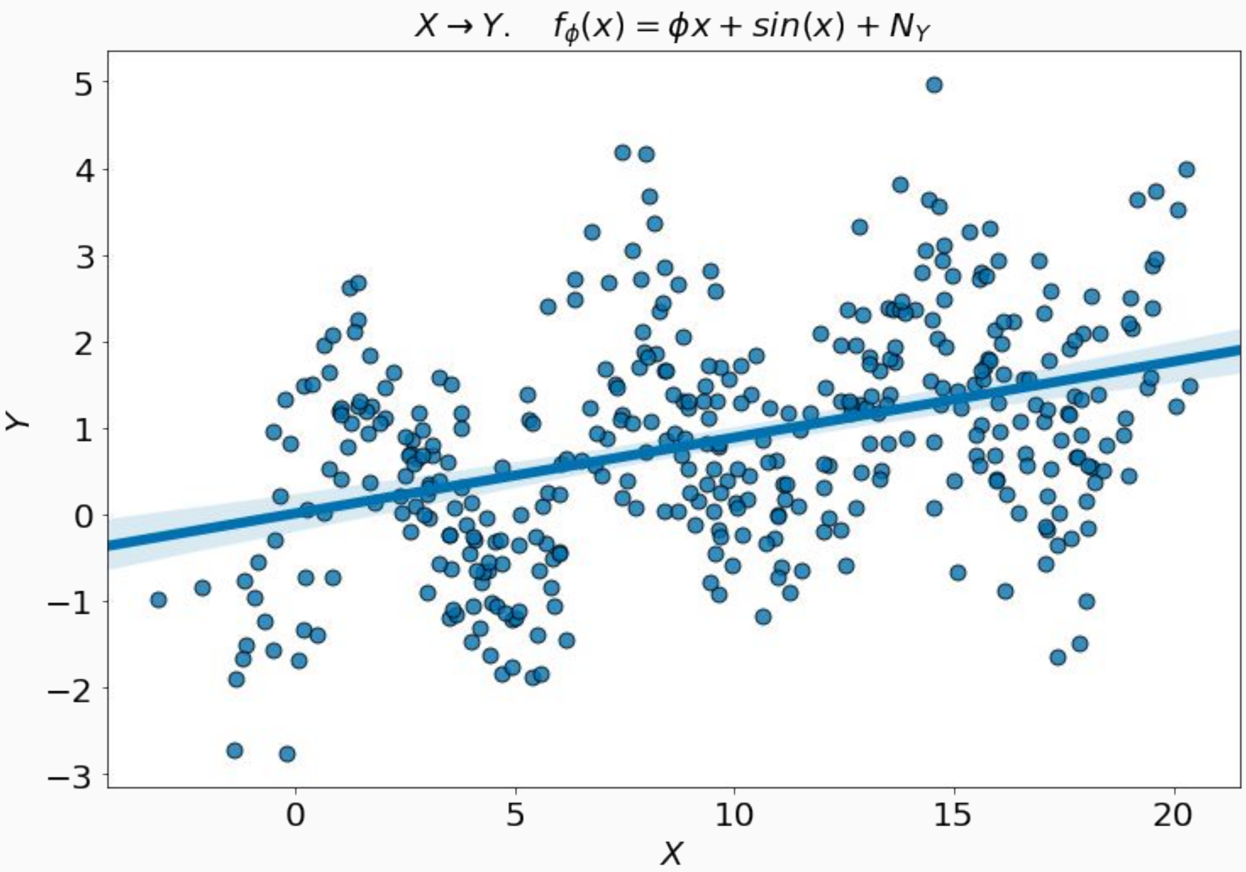


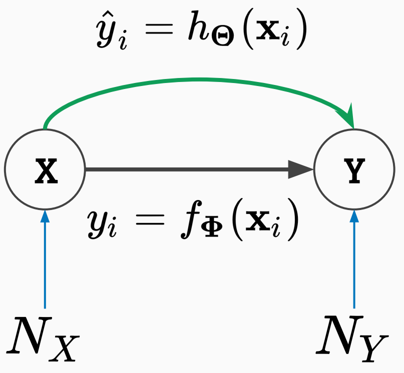
Causalidade e Problemas de Aprendizado. Hipóteses?!



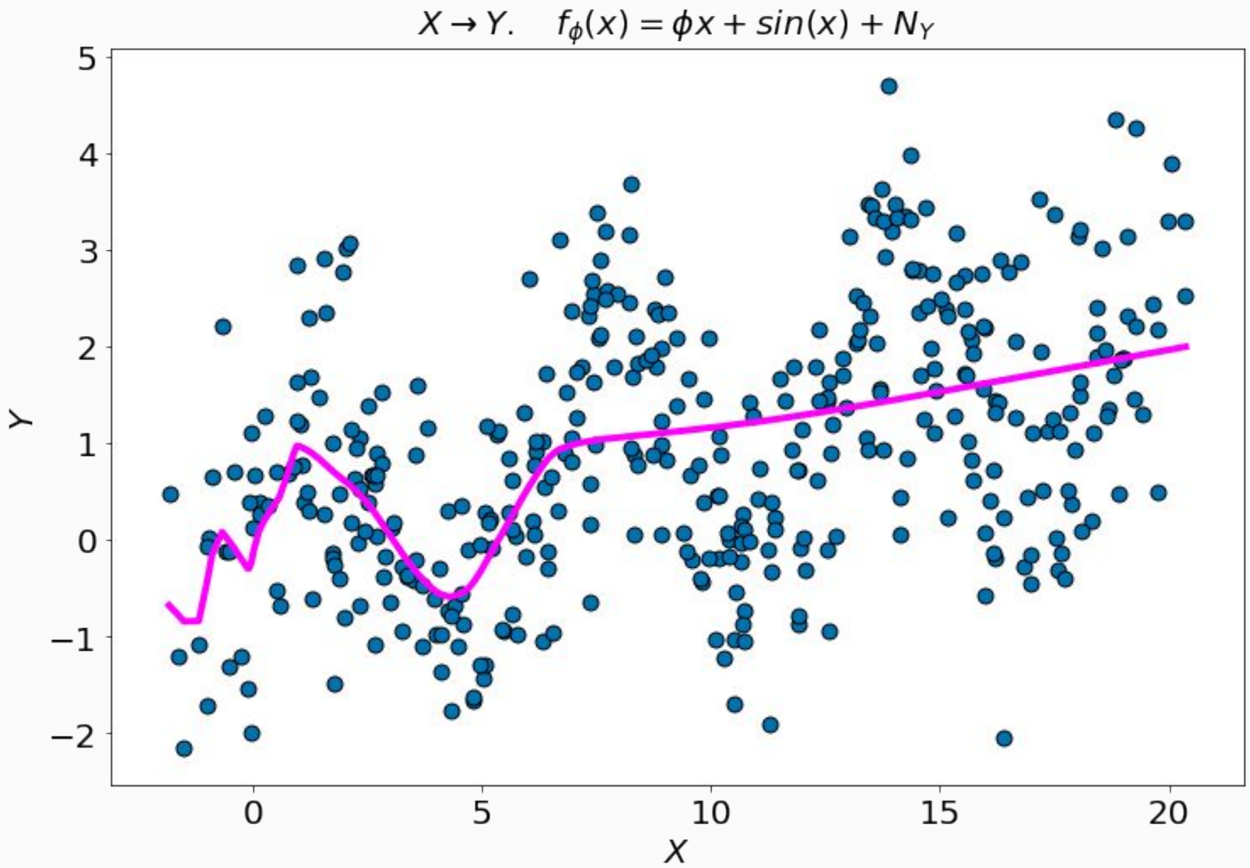


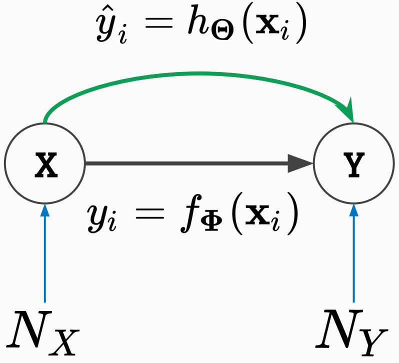
Hipótese Linear





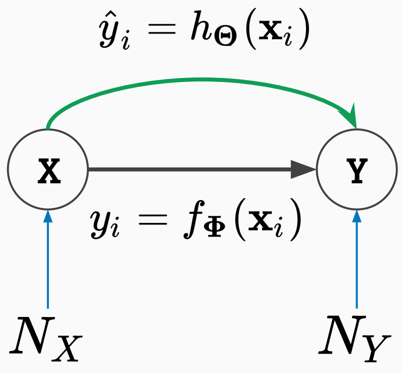
Hipótese Rede Neural 5 camadas de 256 neurons





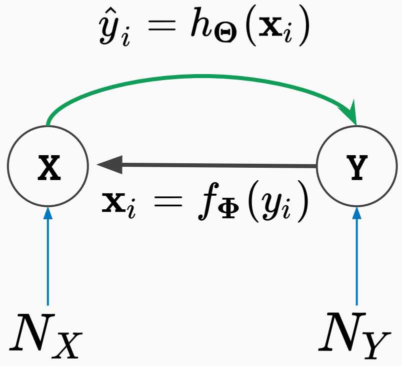
Causalidade e Problemas de Aprendizado

- As hipóteses que testamos até são boas do ponto de vista de previsões. Capturam uma tendências
- Não recuperam o modelo real!
- Tal exemplo mostra a diferença entre a hipótese e o modelo causal. Estamos no mundo de hipóteses com aprendizado profundo, não falamos nada do processo real!

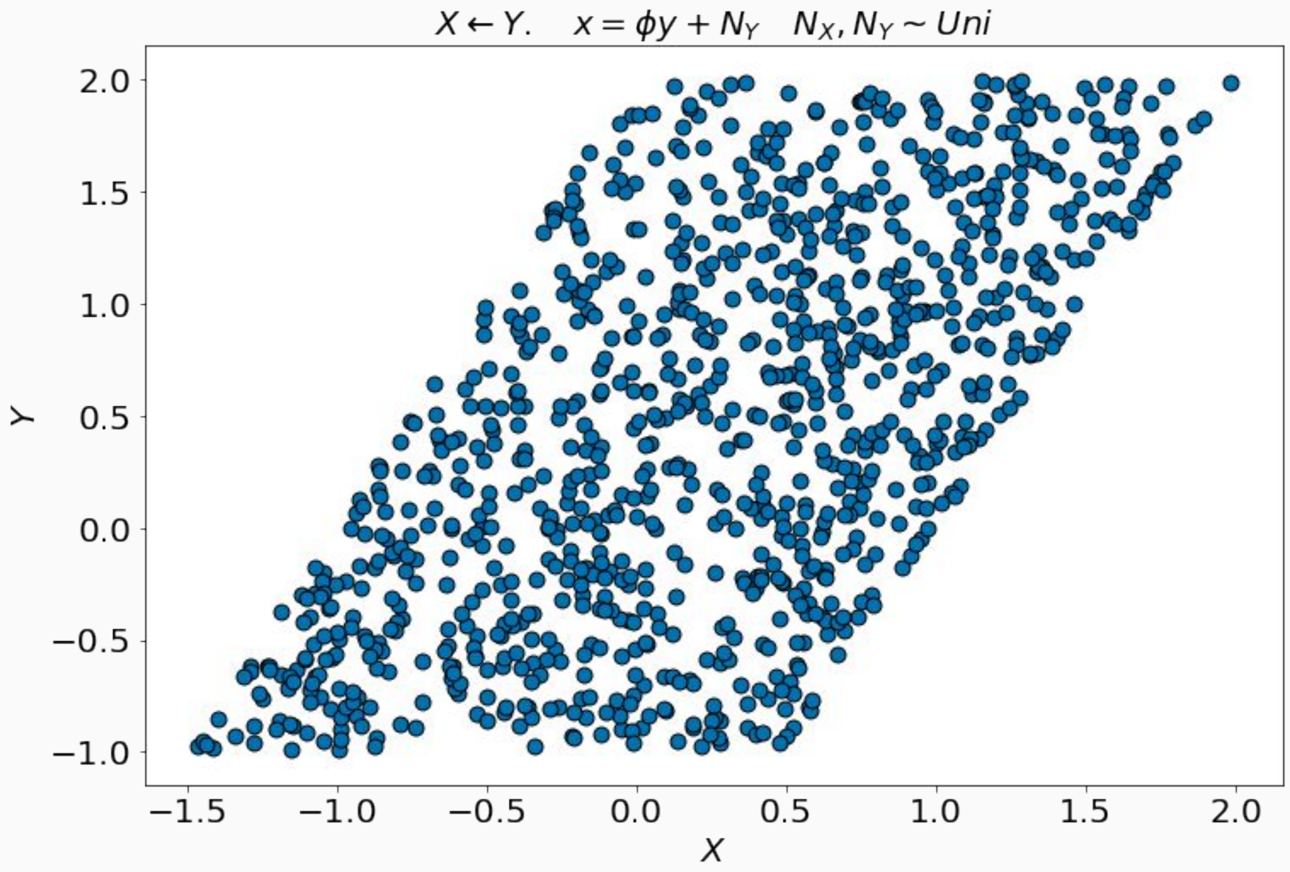


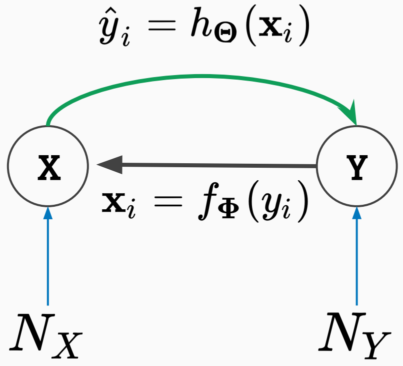
Pensando em Anti-Causalidade

- Vamos inverter a relação e considerar um exemplo
- Aqui os ruídos não serão normais
- Nx e Ny serão uniformes
- Quebra algumas premissas da regressão linear simples!



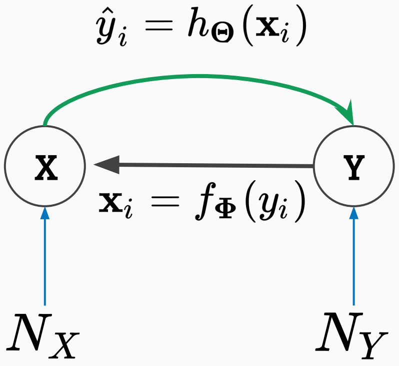
Invertendo a Direção



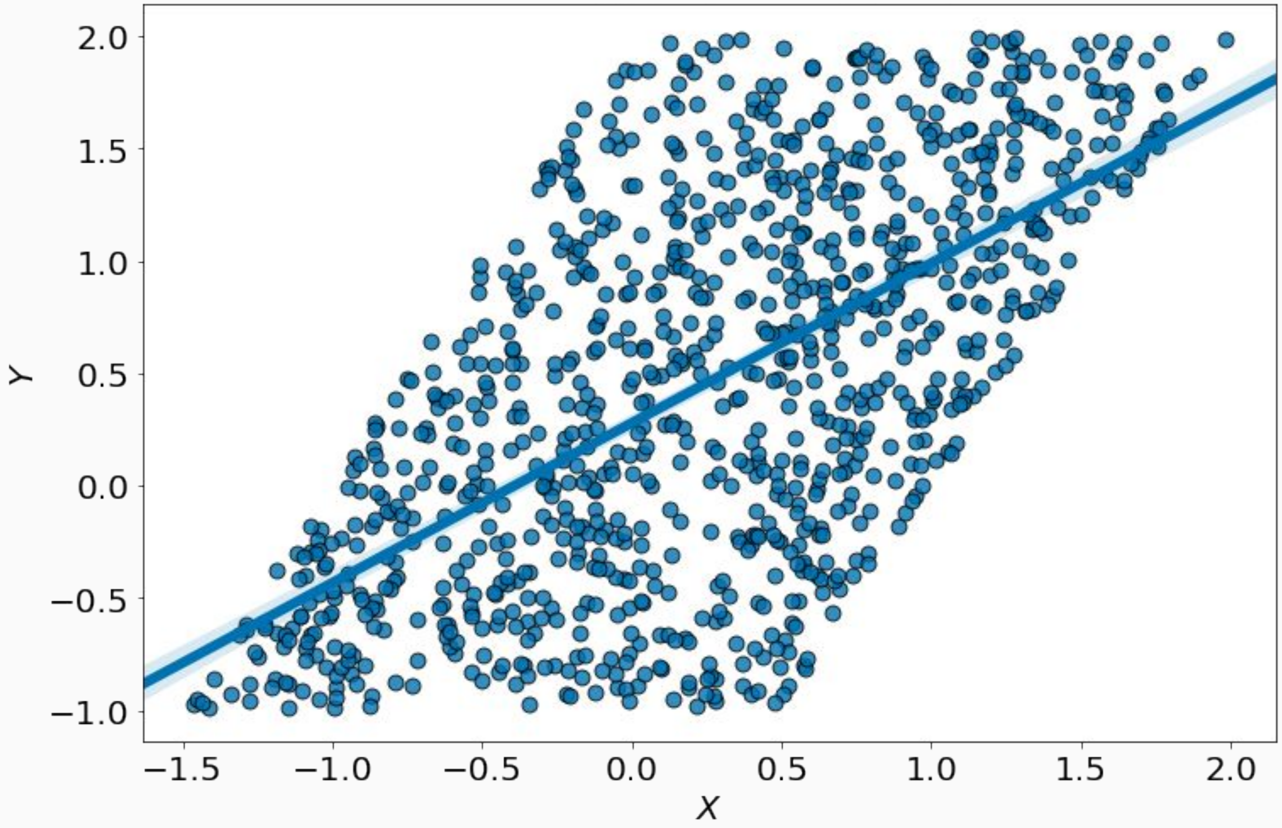


Pensando em Anti-Causalidade

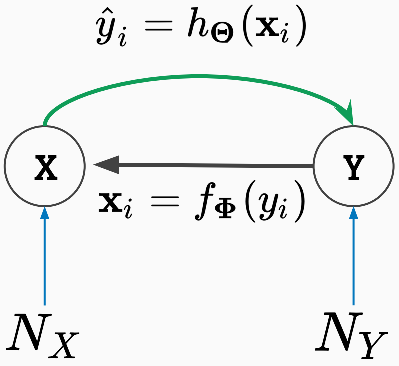
- Vamos tentar recuperar a relação com dois modelos
- Um no sentido anti-causal e outro no causal
- Começaremos do anti-causal



Modelo

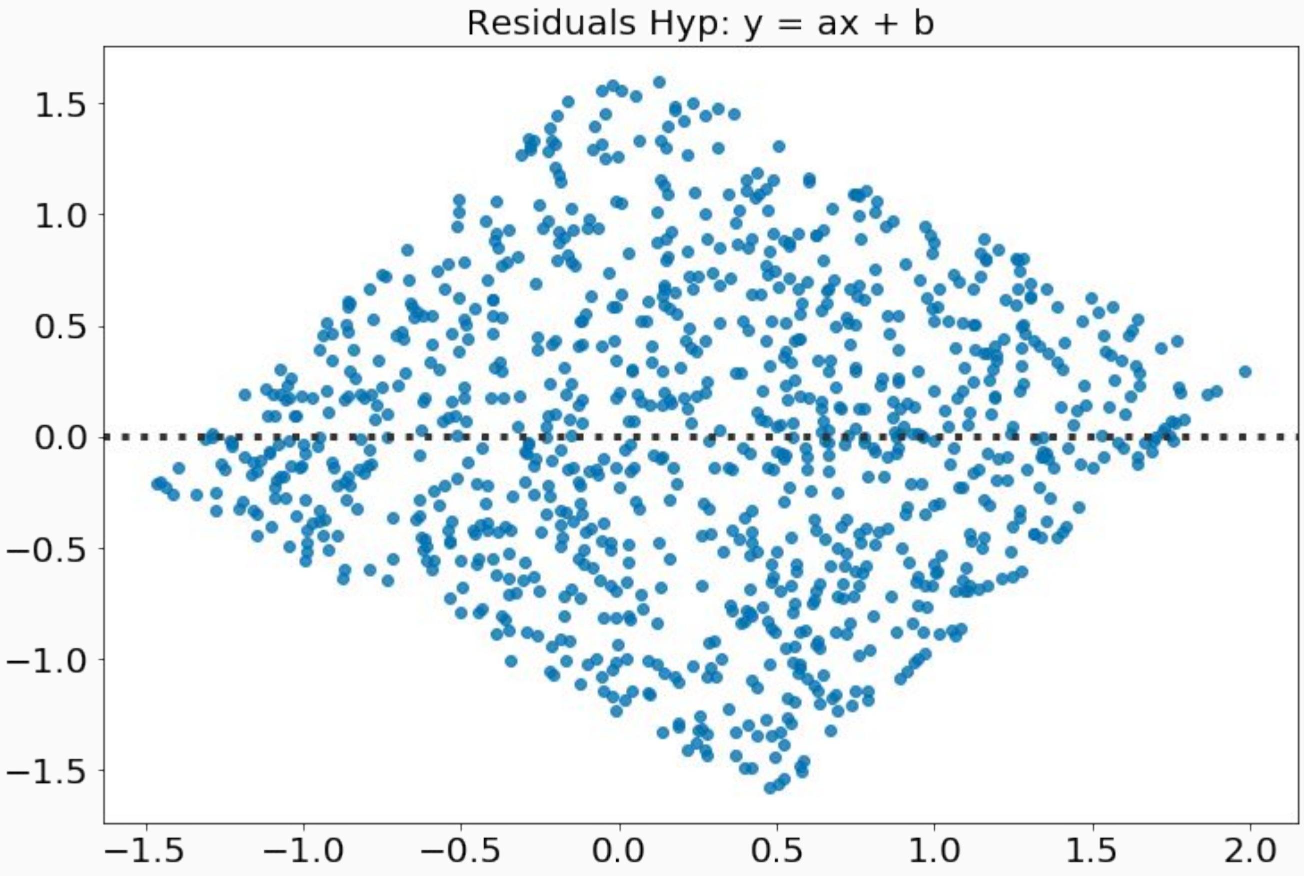


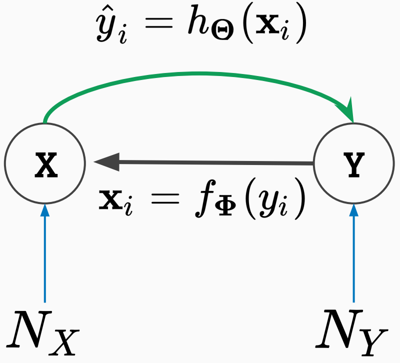
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
Y
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
X



$\hat{y}_i = h_{\mathbf{\Theta}}(\mathbf{x}_i)$
X
Y
$\mathbf{x}_i = f_{\mathbf{\Phi}}(y_i)$
N_X
N_Y

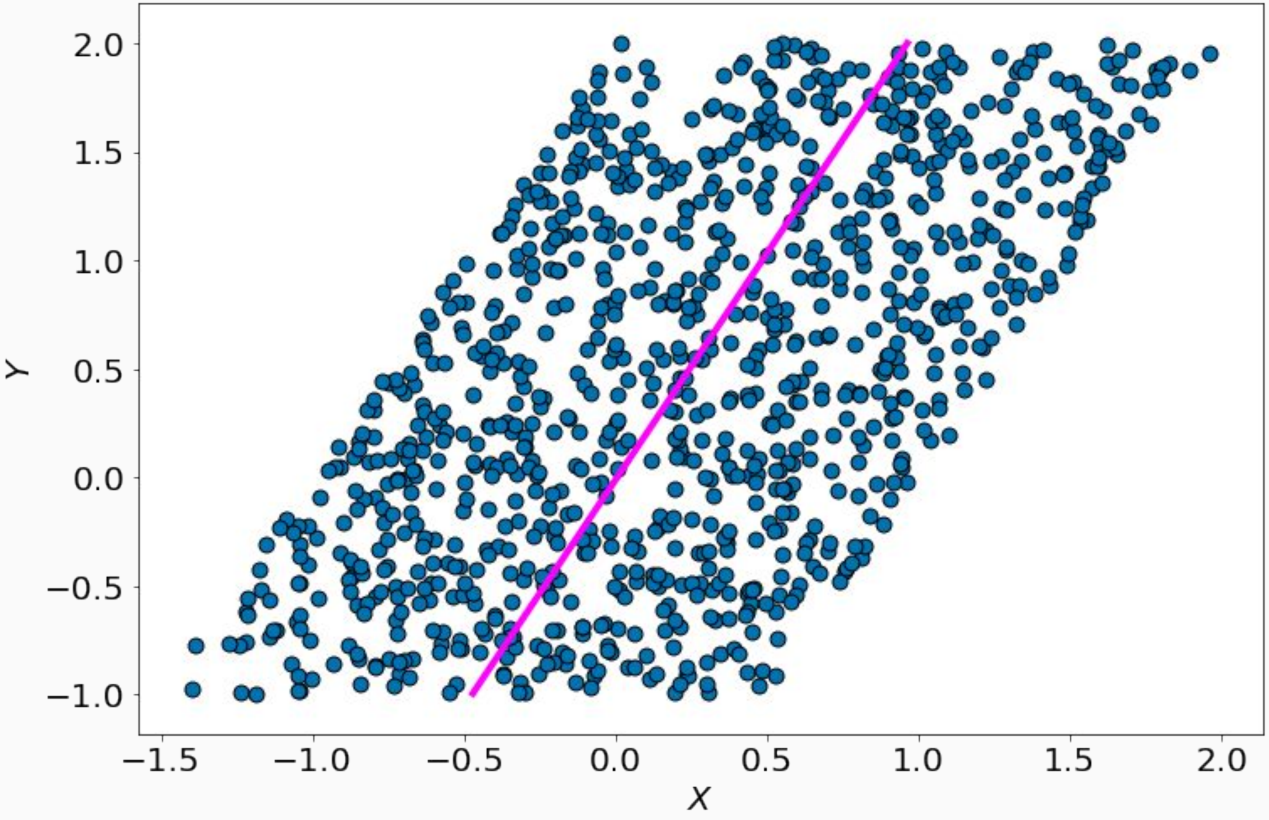
Note como a regressão falha no Residual Plot

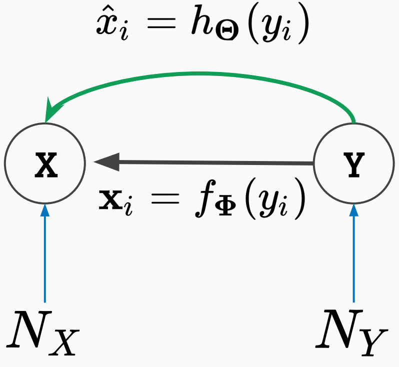




Idealmente,
os erros seriam
uniformes para os
dois lados!

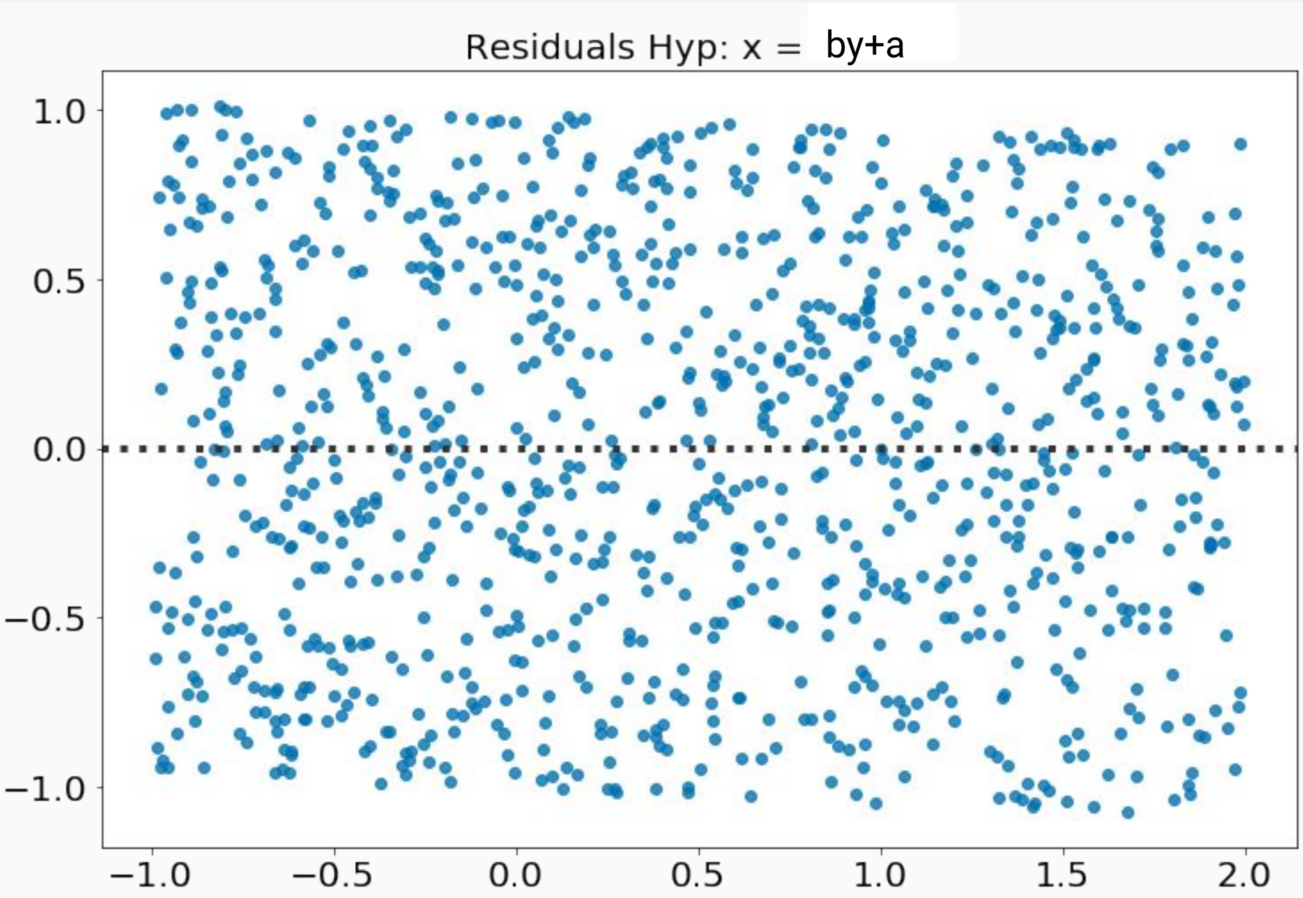
Agora invertendo a previsão! x = by + a

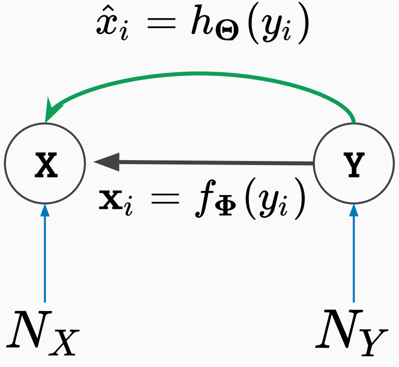




Ao inverter
a relação, bem
melhor!

Residuais



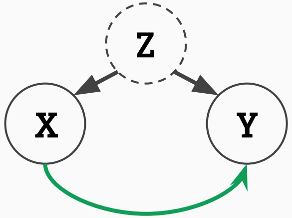


Ao inverter
a relação, bem
melhor!

Podemos pensar em diferentes formas como os dados foram gerados. Como vamos discutir, tais formas trazem impactos para nossos algoritmos.

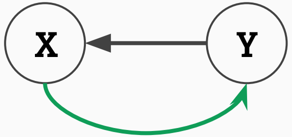
Por clareza, não vamos focar no caso confuso (esquerda).

$$\mathbf{x}_i = f_{1,\mathbf{\Phi}}(z_i)$$
$$y_i = f_{2,\mathbf{\Phi}}(z_i)$$



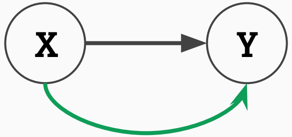
$$\hat{y}_i = h_{\mathbf{\Theta}}(\mathbf{x}_i)$$

$$\mathbf{x}_i = f_{\mathbf{\Phi}}(y_i)$$



$$\hat{y}_i = h_{\mathbf{\Theta}}(\mathbf{x}_i)$$

$$y_i = f_{\mathbf{\Phi}}(\mathbf{x}_i)$$



$$\hat{y}_i = h_{\mathbf{\Theta}}(\mathbf{x}_i)$$

Covariate Shift

Dos Diagramas Anteriores

Observe que temos duas histórias para a mesma equação. Assumindo que y é uma resposta (efeito em alguns livros).

- Focada nos labels, podemos explorar para uma premissa anti-causal

$$p(\mathbf{x}, y) = p(\mathbf{x} \mid y)p(y)$$

- Focada nos dados, podemos explorar para uma premissa causal

$$p(\mathbf{x}, y) = p(y \mid \mathbf{x})p(\mathbf{x})$$

Backtracking

- Da distribuição conjunta você observa um conjunto de dados

$$\mathcal{D} \sim p(\mathbf{x}, y)$$

$$\mathcal{D} = \{(y_i, \mathbf{x}_i)\}$$

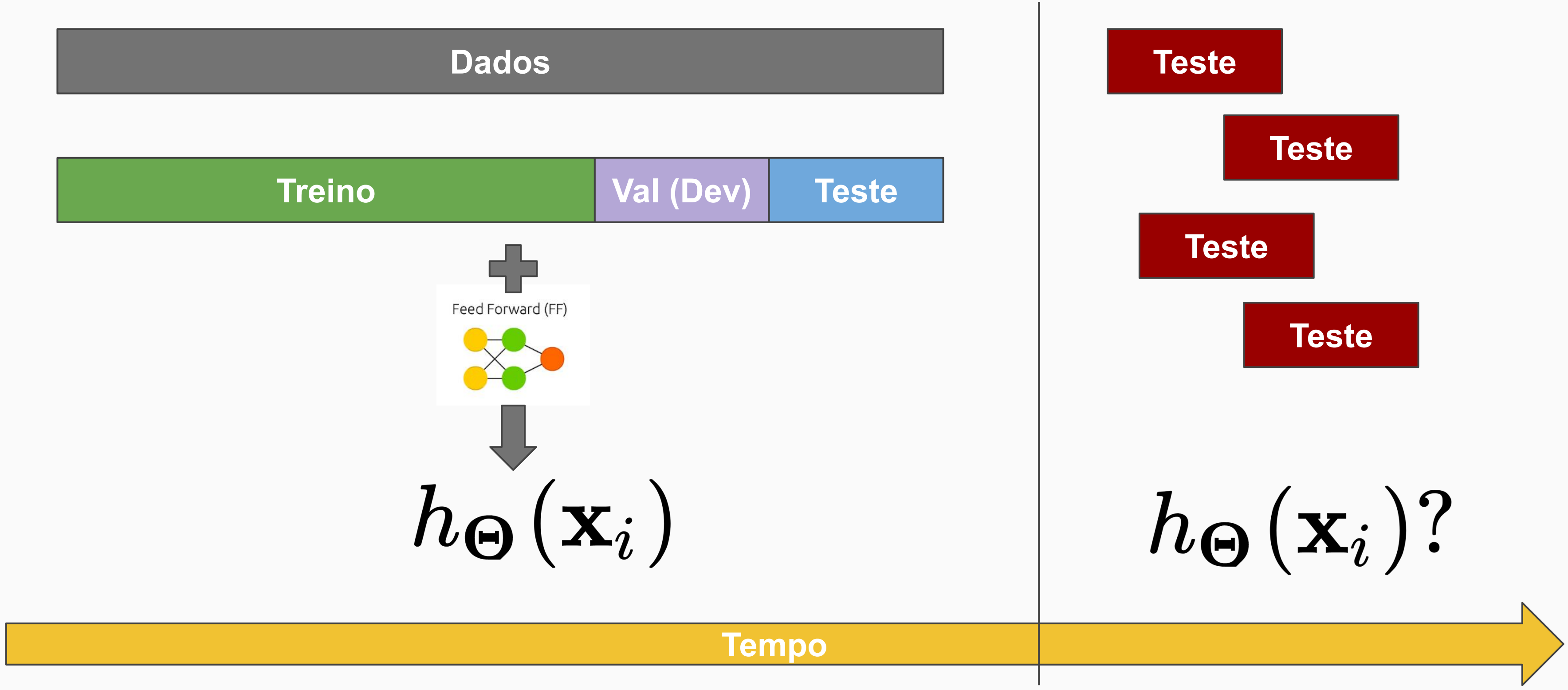
- Usando tais dados, você cria uma hipótese parametrizada

$$\hat{y}_i = h_{\boldsymbol{\Theta}}(\mathbf{x}_i)$$

- Onde a hipótese vem de um modelo. Redes neurais no escopo do nosso curso

A pergunta é: Até quando a hipótese é válida?!

$$\mathcal{D} \sim p(\mathbf{x}, y)$$



Covariate Shift

Pensando no nosso:
Treino vs Teste

- No Treino, observamos

$$p(\mathbf{x}, y) = p(y \mid \mathbf{x})p(\mathbf{x})$$

- No Teste não temos respostas. Porém assumimos que o processo que leva aos dados é o mesmo. Então, podemos assumir que existe uma distribuição.

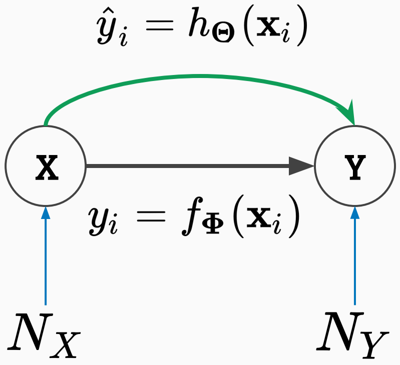
$$q(y \mid \mathbf{x})$$

Não temos certeza da mesma!

Covariate Shift

- Quando estamos no sentido causal
- Podemos assumir que a condicional não muda (nosso f). Então:

$$p(y \mid \mathbf{x}) = q(y \mid \mathbf{x})$$



- No treino

$$p(\mathbf{x}, y) = p(y \mid \mathbf{x})p(\mathbf{x})$$

- Nos testes

$$q(\mathbf{x}, y) = p(y \mid \mathbf{x})q(\mathbf{x})$$

Covariate Shift

- Como o modelo causal não muda, apenas $q(\mathbf{x})$ pode mudar

$$p(y \mid \mathbf{x}) = q(y \mid \mathbf{x})$$

$$p(\mathbf{x}, y) = p(y \mid \mathbf{x})p(\mathbf{x})$$

$$q(\mathbf{x}, y) = p(y \mid \mathbf{x})q(\mathbf{x})$$

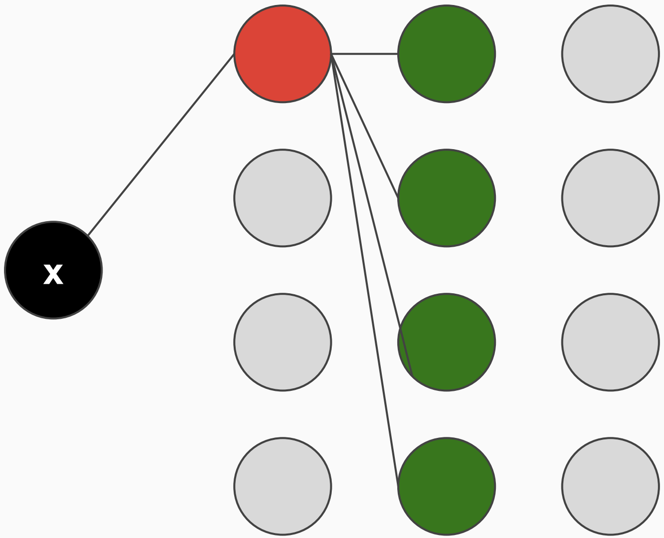
- A nossa hipótese foi invalidada pois os dados mudaram

Interval Covariate Shift

- Covariate shift é relacionado com mudanças em p(x)
- Ao lado temos mudanças no local das vacas
- Porém ao invés de treino e teste ocorre em batches do treino em si
- Isto vai impactar o seu algoritmo de aprendizado!

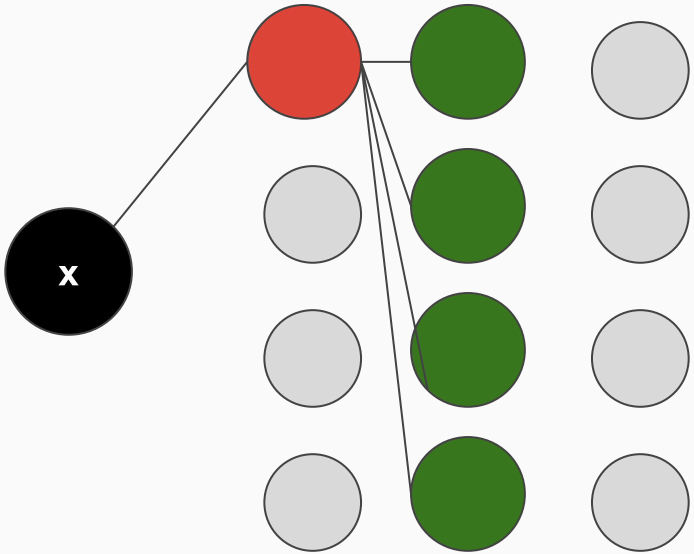
Interval Covariate Shift

- Se os dois conjuntos existem no treino, batch norm ajuda
- Devido ao internal covariate shift



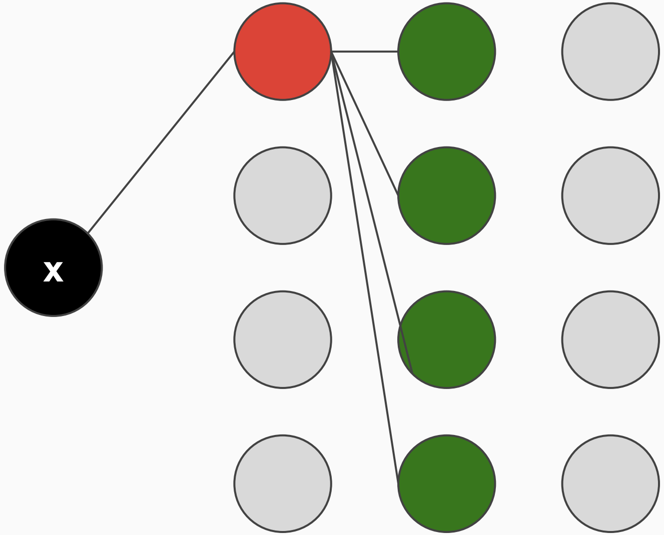
Interval Covariate Shift

- Flutuações em **x** impactam as camadas anteriores
- Como a entrada de uma cada é a saída das outras.



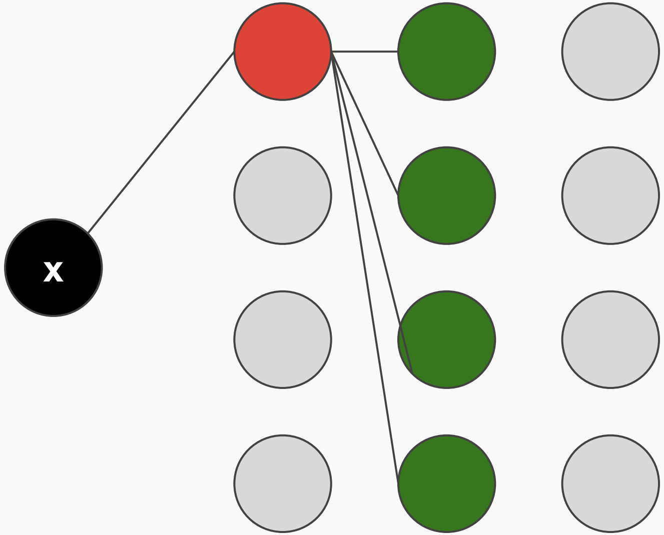
Interval Covariate Shift

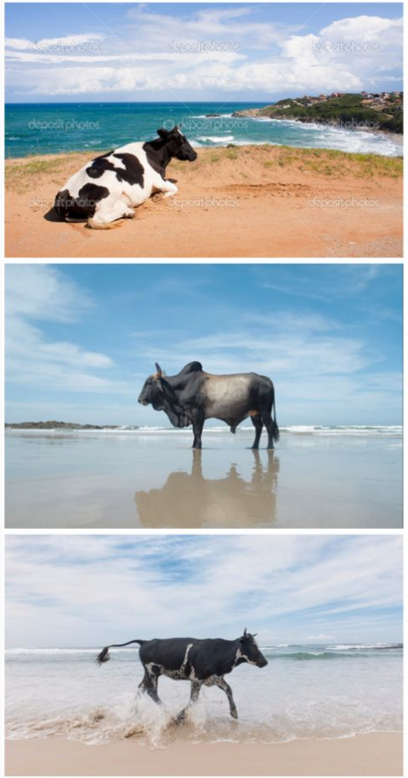
- Internamente, cada camada sofre de **internal** covariate shift!
- As anteriores estão mudando em cada iteração.



Interval Covariate Shift

- Ao fazer batch-norm reduzimos o efeito das mudanças bruscas
- Maior independência entre camadas!





Covariate Shift Real

Mudança ocorre depois do treino. Temos que adaptar:

- Uma abordagem utilizada por alguns autores é estimar: $\mathbb{E}_q[l(\mathbf{x}, y, \theta)]$
- Ou seja, qual o risco no teste (distribuição q)

Covariate Shift Real

Mudança ocorre depois do treino. Temos que adaptar:

- Uma abordagem utilizada por alguns autores é estimar o risco no teste
  $\mathbb{E}_q[l(\mathbf{x}, y, \theta)] = l(y_i, h_\Theta(\mathbf{x_i}))\, q(y_i, \mathbf{x_i})$
- Usando zero-one loss. No treino temos o risco abaixo

$$\mathbb{E}_p[l(\mathbf{x}, y, \theta)] = \sum_i (y_i - h_\Theta(\mathbf{x_i}))\, p(y_i, \mathbf{x_i})$$

$$\mathbb{E}_p[l(\mathbf{x}, y, \theta)] = \sum_i (y_i - h_\Theta(\mathbf{x_i}))\, p(y_i \mid \mathbf{x_i}) p(\mathbf{x_i})$$

Covariate Shift Real

Mudança ocorre depois do treino. Temos que adaptar:

- Uma abordagem utilizada por alguns autores é estimar o risco no teste
  $\mathbb{E}_q[l(\mathbf{x}, y, \theta)] = l(y_i, h_\Theta(\mathbf{x_i}))\, q(y_i, \mathbf{x_i})$
- Usando zero-one loss. No treino temos o risco abaixo

$$\mathbb{E}_p[l(\mathbf{x}, y, \theta)] = \sum_i (y_i - h_\Theta(\mathbf{x_i}))\, p(y_i, \mathbf{x_i})$$
$$\mathbb{E}_p[l(\mathbf{x}, y, \theta)] = \sum_i (y_i - h_\Theta(\mathbf{x_i}))\, p(y_i \mid \mathbf{x_i}) p(\mathbf{x_i})$$

- No teste teremos (lembrando da premissa que $p(y \mid \mathbf{x}) = q(y \mid \mathbf{x})$)

$$\mathbb{E}_q[l(\mathbf{x}, y, \theta)] = \sum_i (y_i - h_\Theta(\mathbf{x_i}))\, q(y_i, \mathbf{x_i})$$
$$\mathbb{E}_q[l(\mathbf{x}, y, \theta)] = \sum_i (y_i - h_\Theta(\mathbf{x_i}))\, p(y_i \mid \mathbf{x_i}) q(\mathbf{x_i})$$

- Não precisa ser zero-one, pode ser erro quadrado ou qualquer coisa!

Covariate Shift Real

Pensando em um novo batch de teste, podemos re-escrever o risco no teste como uma função do risco no treino. Basta multiplicar e dividir pela densidade no treino.

$$\mathbb{E}_q[l(\mathbf{x}, y, \theta)] = \sum_i l(y_i, h_\Theta(\mathbf{x_i}))\, q(y_i, \mathbf{x_i})$$

$$\mathbb{E}_q[l(\mathbf{x}, y, \theta)] = \sum_i l(y_i, h_\Theta(\mathbf{x_i}))\, p(y_i \mid \mathbf{x_i}) q(\mathbf{x_i})$$

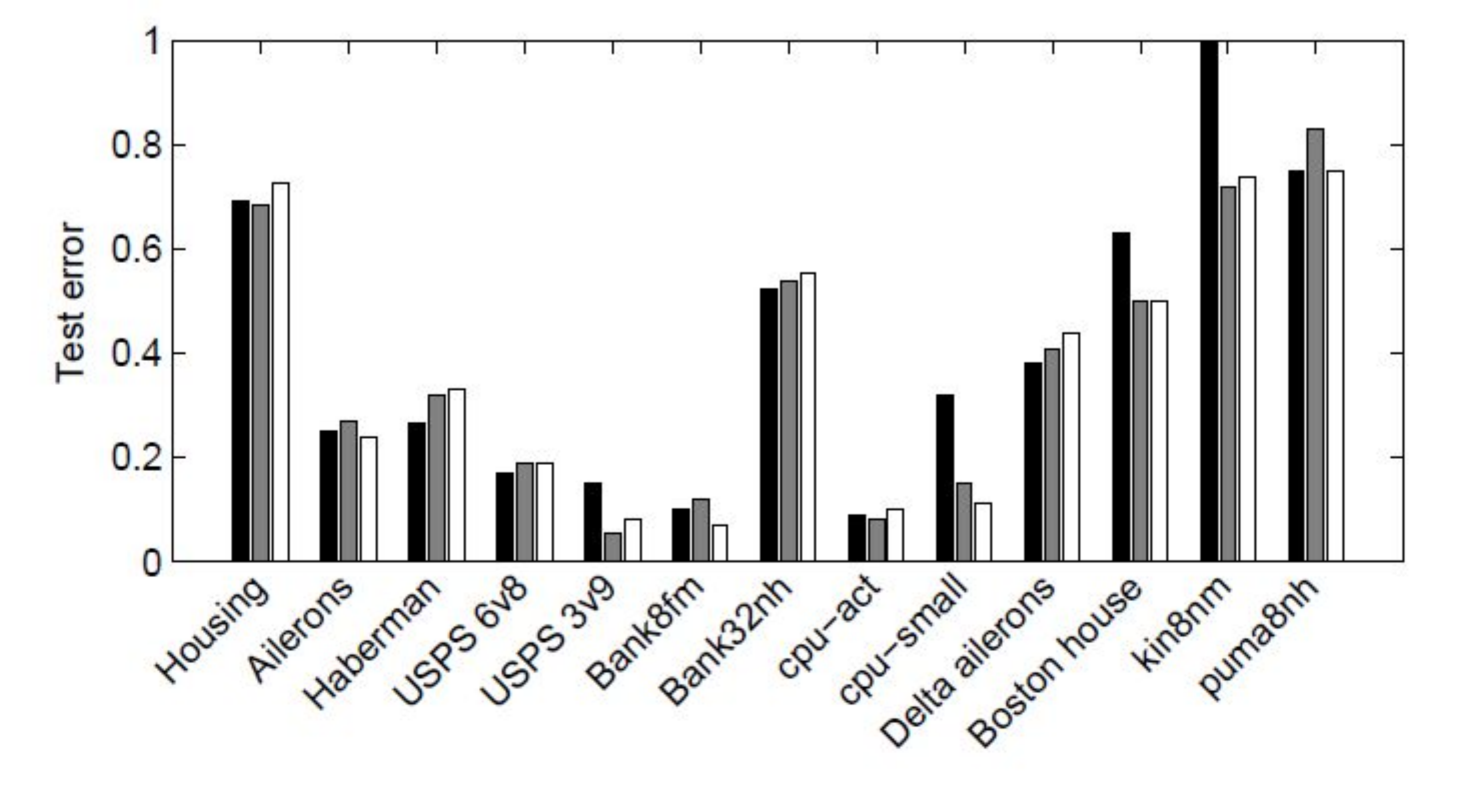
$$\mathbb{E}_q[l(\mathbf{x}, y, \theta)] = \sum_i l(y_i, h_\Theta(\mathbf{x_i}))\, p(y_i \mid \mathbf{x_i}) q(\mathbf{x_i}) \frac{p(\mathbf{x_i})}{p(\mathbf{x_i})}$$

$$\mathbb{E}_q[l(\mathbf{x}, y, \theta)] = \mathbb{E}_p[l(\mathbf{x}, y, \theta)] \frac{q(\mathbf{x})}{p(\mathbf{x})}$$

Agora

- Basta estimar q(**x**)/p(**x**)
- Usando Densidades + Rejection Sampling
  Zadrozny 2003 e 2004
  http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.170&rep=rep1&type=pdf
- Usando Kernel Means Matching
  http://www.gatsby.ucl.ac.uk/~gretton/papers/covariateShiftChapter.pdf

Resultados


1
0.8
0.6
0.4
0.2
0
Test error
Housing
Ailerons
Haberman
USPS 6v8
USPS 3v9
Bank8fm
Bank32nh
cpu-act
cpu-small
Delta ailerons
Boston house
kin8nm
puma8nh

Label Shift

Label Shift

- Ao invés de mudanças nas features, número de instâncias nas classes mudam
- Mais pneumonia no inverno
- Mais bots na época de eleições
- Mais chance de achar petróleo em um local em comparação com outro

Mesma Ideia

- Porém agora, vamos partir de uma premissa anti-causal (mais comum)

$$p(\mathbf{x}, y) = p(\mathbf{x} \mid y)p(y)$$

- Assumindo que no teste

$$q(\mathbf{x} \mid y) = p(\mathbf{x} \mid y)$$

- Aqui, se os dados mudam eles mudam por conta de p(y)

Mesma Ideia

- Podemos chegar no mesmo resultado de antes trocando x por y.

$$\mathbb{E}_q[l(\mathbf{x}, y, \theta)] = \mathbb{E}_p[l(\mathbf{x}, y, \theta)]\,\frac{q(y)}{p(y)}$$

- O interessante é que:
    - A premissa anti-causal é mais comum nas nossas tarefas
    - Precisamos estimar distribuição de rótulos/labels

Estimando Q e P

- Um resultado bem interessante [Lipton 2018] é que podemos estimar q e p usando um classificador qualquer (black-box)
- Precisamos apenas da matriz de confusão do classificador

A.1 The *label shift* (also known as *target shift*) assumption

$$p(\boldsymbol{x}|y) = q(\boldsymbol{x}|y) \quad \forall\, x \in \mathcal{X},\; y \in \mathcal{Y}.$$

A.2 For every $y \in \mathcal{Y}$ with $q(y) > 0$ we require $p(y) > 0$.[2]

A.3 Access to a black box predictor $f : \mathcal{X} \to \mathcal{Y}$ where the expected confusion matrix $\mathbf{C}_p(f)$ is invertible.

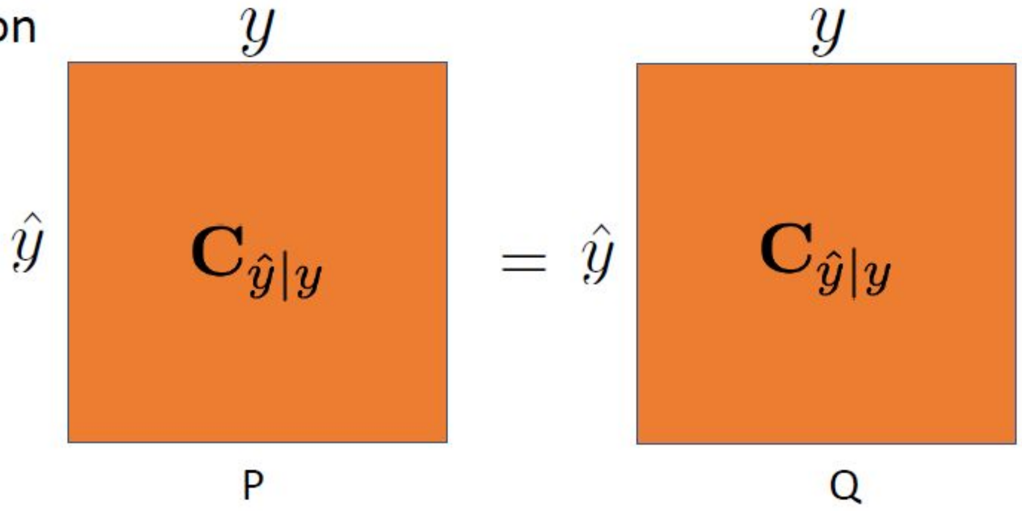
$$\mathbf{C}_P(f) := p(f(\boldsymbol{x}), y) \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{Y}|}$$

Slide de Zachary Lipton

Applying the label shift assumption...

- $\mathbf{C}_{\hat{y}|y}$ - column-normalized is **identical** in under P and Q

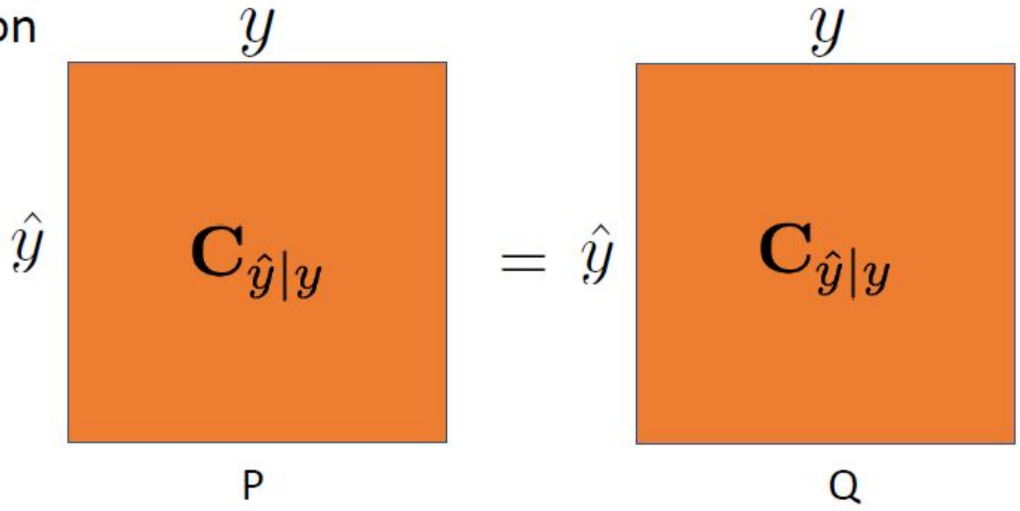
- We can estimate confusion matric on P
- Don't need to observe labels from Q



O truque para entender é focar no column-normalized na assumption anti-causal

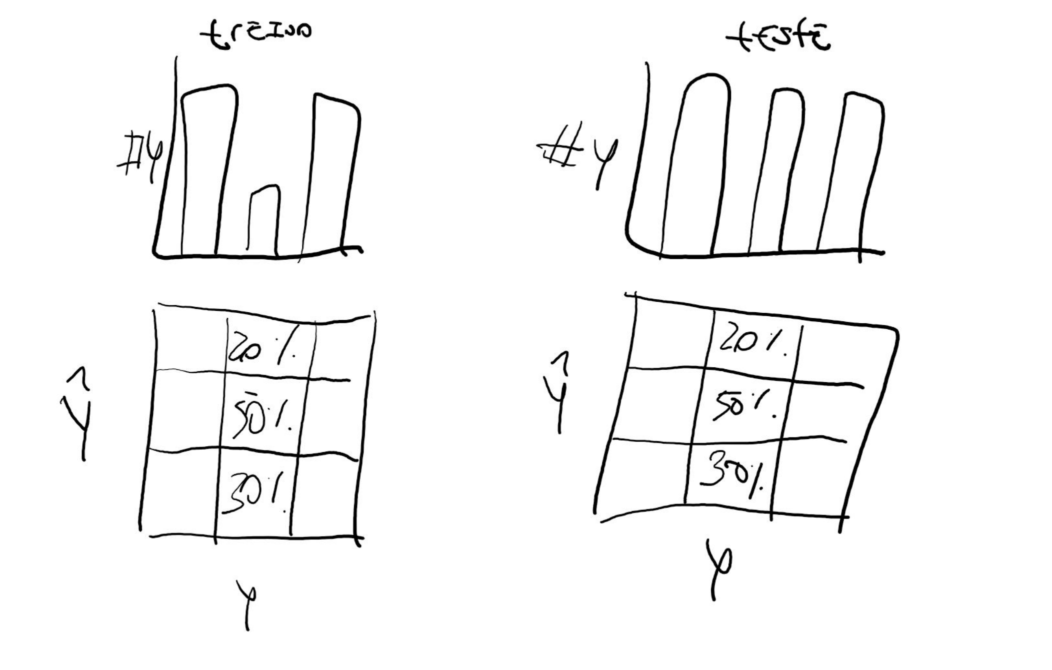
Applying the label shift assumption...

- $\mathbf{C}_{\hat{y}|y}$ - column-normalized is **identical** in under P and Q
- We can estimate confusion matric on P
- Don't need to observe labels from Q



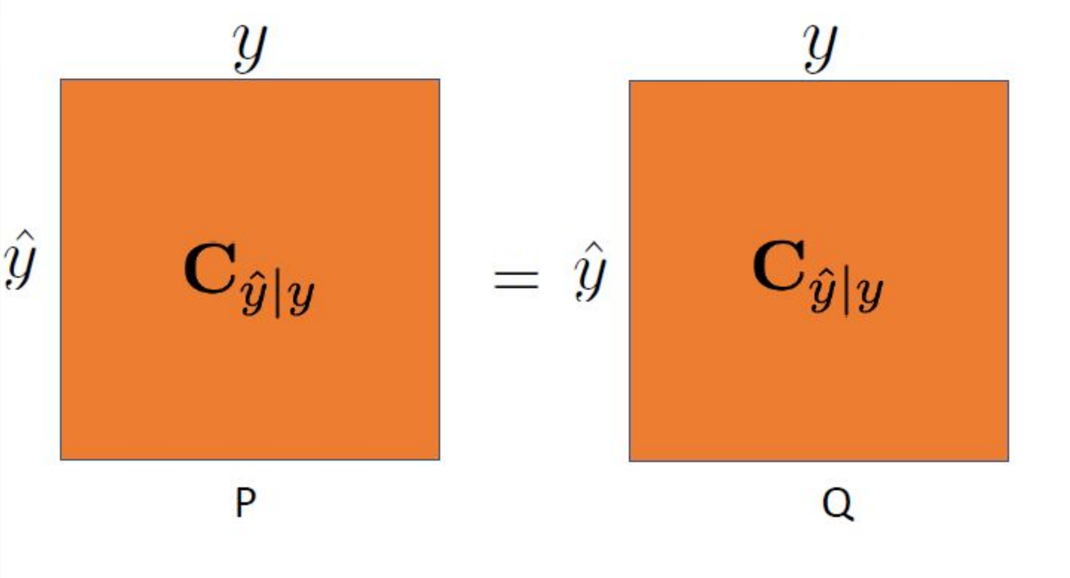
O truque para entender é focar no column-normalized na assumption anti-causal

- Se $p(\mathbf{x}|y) = q(\mathbf{x}|y)$
- Mesmo com um aumento de labels…
- Teremos as mesmas features para cada instância



Portanto

- Podemos estimar a matriz de confusão usando dados do treino
- A mesma ainda é válida no teste
- Agora podemos estimar q(y). p(y) já vem do treino.



Portanto

- Aplicando a regra de bayes podemos recuperar q(y) no teste
- p(y) já temos diretamente do treino
- Resolvendo abaixo

$$
\begin{aligned}
q(\hat{y}) &= \sum_{y \in \mathcal{Y}} q(\hat{y}|y) q(y) \\
&= \sum_{y \in \mathcal{Y}} p(\hat{y}|y) q(y) = \sum_{y \in \mathcal{Y}} p(\hat{y}, y) \frac{q(y)}{p(y)}
\end{aligned}
$$

Portanto

- Aplicando a regra de bayes podemos recuperar q(y) no teste
- p(y) já temos diretamente do treino
- Resolvendo abaixo já temos tudo!

PREVISÕES

TREINO

$$q(\hat{y}) = \sum_{y \in \mathcal{Y}} \underline{\underline{q(\hat{y}|y)}}\, q(y)$$

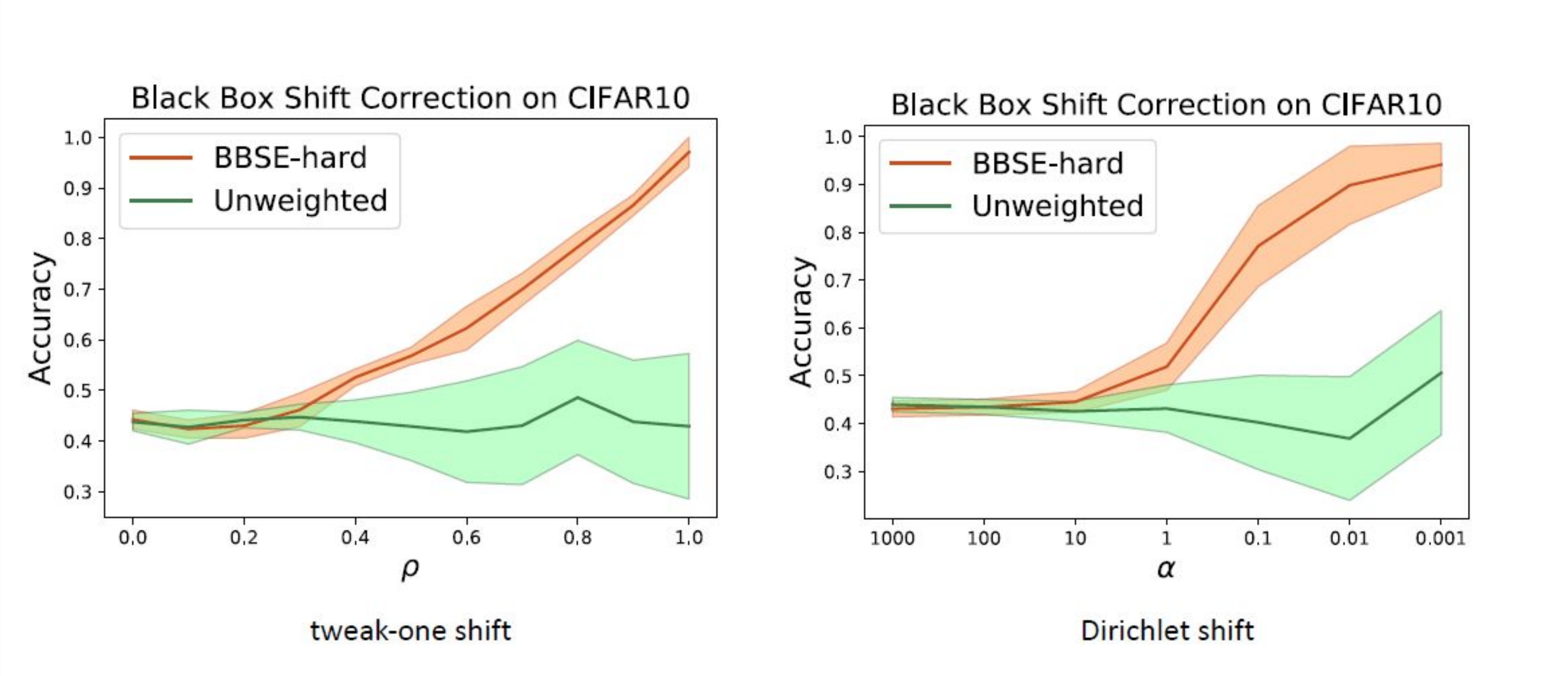
$$= \sum_{y \in \mathcal{Y}} \underline{\underline{p(\hat{y}|y)}}\, q(y) = \sum_{y \in \mathcal{Y}} p(\hat{y}, y) \frac{q(y)}{p(y)}$$

Solve q(y)

CONFUSION

Agora Podemos Treinar

- Para cada novo batch, melhorar o classificador usando q/p
- $\mathbb{E}_q[l(\mathbf{x}, y, \theta)] = \mathbb{E}_p[l(\mathbf{x}, y, \theta)]\frac{q(y)}{p(y)}$

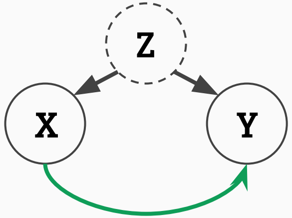


Conclusões

Shifts em Dados

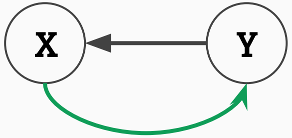
- É mais comum do que é discutido
- Mais simples assumir que nada muda
- Em casos particulares, podemos fazer alguma coisa
- Agora, se o modelo causal muda :-(
    - Funções f abaixo

$$\mathbf{x}_i = f_{1,\mathbf{\Phi}}(z_i)$$
$$y_i = f_{2,\mathbf{\Phi}}(z_i)$$



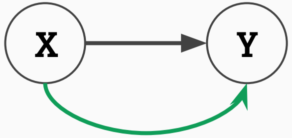
$$\hat{y}_i = h_{\mathbf{\Theta}}(\mathbf{x}_i)$$

$$\mathbf{x}_i = f_{\mathbf{\Phi}}(y_i)$$



$$\hat{y}_i = h_{\mathbf{\Theta}}(\mathbf{x}_i)$$

$$y_i = f_{\mathbf{\Phi}}(\mathbf{x}_i)$$



$$\hat{y}_i = h_{\mathbf{\Theta}}(\mathbf{x}_i)$$

Causalidade e Adaptação

- Assumindo que o modelo causal não muda
  - Podemos nos adaptar
  - Ponderando novos batches leva para classificadores melhores
- Trabalhar no mundo anti-causal é normal
  - Label-shift é fácil de implementar
- Problemas em aberto:
  - Dados que não vêm em batches
    - Um por vez
  - Aplicar as ideias para shifts mais extremos
    - Domain Shift
  - Outras formas de estimar q(x) e p(x)
    - GANs

Referências

(1) Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift
Stephan Rabanser, Stephan Gunnemann, Zachary C. Lipton - ICML 2019
https://arxiv.org/pdf/1810.11953.pdf

(2) Bernhard Schölkopf: Learning Independent Mechanisms
https://sites.google.com/view/nips2018causallearning/home

(3) Trustworthy Deep Learning
https://berkeley-deep-learning.github.io/cs294-131-s19/