Ensembles II

Jiayu Zhou

¹Department of Computer Science and Engineering Michigan State University East Lansing, MI USA

March 17, 2016

Slides from "A Gental Introduction to Gradient Boosting." by Cheng Li.

Table of contents

- Introduction
- 2 Regression
 - Choice of Loss Functions
- Classification
 - Letter Recognition
 - Kullback-Leibler (KL) divergence
 - Gradient Boosting for Classification

Introduction

Gradient Boosting

- A powerful machine learning algorithm
- Regression/Classification/Ranking.
- Won Track 1 of the Yahoo Learning to Rank Challenge

- Gradient Boosting = Gradient Descent + Boosting
- AdaBoost

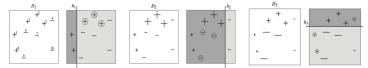


Figure: AdaBoost. Source: Figure 1.1 of [Schapire and Freund, 2012]

- Gradient Boosting = Gradient Descent + Boosting
- AdaBoost

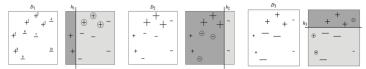


Figure: AdaBoost. Source: Figure 1.1 of [Schapire and Freund, 2012]

- Fit an additive model (ensemble) $\sum_t \rho_t h_t(x)$ in a forward stage-wise manner.
 - In each stage, introduce a weak learner to compensate the shortcomings of existing weak learners.
 - In Adaboost, "shortcomings" are identified by high-weight data points.

- Gradient Boosting = Gradient Descent + Boosting
- AdaBoost

$$H(x) = \sum_{t} \rho_t h_t(x)$$

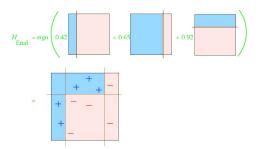


Figure: AdaBoost. Source: Figure 1.1 of [Schapire and Freund, 2012]

- Gradient Boosting = Gradient Descent + Boosting
- Gradient Boosting
 - Fit an additive model (ensemble) $\sum_t \rho_t h_t(x)$ in a forward stage-wise manner.
 - In each stage, introduce a weak learner to compensate the shortcomings of existing weak learners.
 - In Gradient Boosting, shortcomings are identified by gradients.
 - In Adaboost, shortcomings are identified by high-weight data points.
 - Both high-weight data points and gradients tell us how to improve our model.

Regression

Lets play a game...

You are given $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, and the task is to fit a model F(x) to minimize square loss.

Suppose your friend wants to help you and gives you a model F. You check his model and find the model is good but not perfect. There are some mistakes: $F(x_1)=0.8$, while $y_1=0.9$, and $F(x_2)=1.4$ while $y_2=1.3$... How can you improve this model?

Rule of the game:

- You are not allowed to remove anything from F or change any parameter in F.
- You can add an additional model (regression) h to F, so the new prediction will be F(x) + h(x).

Simple solution: You wish to improve the model such that

$$F(x_1) + h(x_1) = y_1$$

$$F(x_2) + h(x_2) = y_2$$

$$\dots$$

$$F(x_n) + h(x_n) = y_n$$

Simple solution: Or, equivalently, you wish

$$h(x_1) = y_1 - F(x_1)$$

$$h(x_2) = y_2 - F(x_2)$$

$$\dots$$

$$h(x_n) = y_n - F(x_n)$$

Can any regression model achieve goal perfectly?

Simple solution: Or, equivalently, you wish

$$h(x_1) = y_1 - F(x_1)$$

$$h(x_2) = y_2 - F(x_2)$$

$$\dots$$

$$h(x_n) = y_n - F(x_n)$$

Can any regression model achieve goal perfectly? Maybe not ...

Simple solution: Or, equivalently, you wish

$$h(x_1) = y_1 - F(x_1)$$

$$h(x_2) = y_2 - F(x_2)$$

$$\dots$$

$$h(x_n) = y_n - F(x_n)$$

Can any regression model achieve goal perfectly?

Maybe not ...

But some regression models might be able to do this approximately.

Simple solution: Or, equivalently, you wish

$$h(x_1) = y_1 - F(x_1)$$

$$h(x_2) = y_2 - F(x_2)$$

$$\dots$$

$$h(x_n) = y_n - F(x_n)$$

Can any regression model achieve goal perfectly?

Maybe not ...

But some regression models might be able to do this approximately.

How?

Simple solution: Or, equivalently, you wish

$$h(x_1) = y_1 - F(x_1)$$

$$h(x_2) = y_2 - F(x_2)$$

$$\dots$$

$$h(x_n) = y_n - F(x_n)$$

Can any regression model achieve goal perfectly?

Maybe not ...

But some regression models might be able to do this approximately.

How?

Just fit a regression model h to data

$$(x_1, y_1 - F(x_1)), (x_2, y_2 - F(x_2)), ..., (x_n, y_n - F(x_n))$$

Simple solution: Or, equivalently, you wish

$$h(x_1) = y_1 - F(x_1)$$

$$h(x_2) = y_2 - F(x_2)$$

$$\dots$$

$$h(x_n) = y_n - F(x_n)$$

Can any regression model achieve goal perfectly?

Maybe not ...

But some regression models might be able to do this approximately.

How?

Just fit a regression model h to data

$$(x_1, y_1 - F(x_1)), (x_2, y_2 - F(x_2)), ..., (x_n, y_n - F(x_n))$$

Congratulations, you get a better model!

Simple solution:

 $y_i - F(x_i)$ are called residuals. These are the parts that existing model F cannot do well.

The role of h is to compensate the shortcoming of existing model F.

12 / 58

Simple solution:

 $y_i - F(x_i)$ are called residuals. These are the parts that existing model F cannot do well.

The role of h is to compensate the shortcoming of existing model F. If the new model F+h is still not satisfactory,

Simple solution:

 $y_i - F(x_i)$ are called residuals. These are the parts that existing model F cannot do well.

The role of h is to compensate the shortcoming of existing model F. If the new model F+h is still not satisfactory, we can add another regression model...

Simple solution:

 $y_i - F(x_i)$ are called residuals. These are the parts that existing model F cannot do well.

The role of h is to compensate the shortcoming of existing model F. If the new model F+h is still not satisfactory, we can add another regression model...

We are improving the predictions of training data, is the procedure also useful for test data?

Simple solution:

 $y_i - F(x_i)$ are called residuals. These are the parts that existing model F cannot do well.

The role of h is to compensate the shortcoming of existing model F. If the new model F+h is still not satisfactory, we can add another regression model...

We are improving the predictions of training data, is the procedure also useful for test data?

Yes! Because we are building a model, and the model can be applied to test data as well.

Simple solution:

 $y_i - F(x_i)$ are called residuals. These are the parts that existing model F cannot do well.

The role of h is to compensate the shortcoming of existing model F. If the new model F+h is still not satisfactory, we can add another regression model...

We are improving the predictions of training data, is the procedure also useful for test data?

Yes! Because we are building a model, and the model can be applied to test data as well.

How is this related to gradient descent?

Gradient Descent: Minimize a function by moving in the opposite direction of the gradient.

$$\theta_i = \theta_i - \rho \frac{\partial J}{\partial \theta_i}, \quad \boldsymbol{\theta} = \boldsymbol{\theta} - \rho \nabla_{\boldsymbol{\theta}} J$$

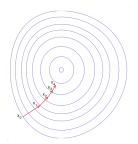


Figure: Gradient Descent.

Source: http://en.wikipedia.org/wiki/Gradient_descent

How is this related to gradient descent?

Loss function $L(y, F(x)) = (y - F(x))^2/2$

We want to minimize $J = \sum_i L(y_i, F(x_i))$ by adjusting

$$F(x_1), F(x_2), \ldots, F(x_n).$$

Notice that $F(x_1), F(x_2), \dots, F(x_n)$ are just some numbers. We can treat $F(x_i)$ as parameters and take derivatives.

How is this related to gradient descent?

Loss function $L(y, F(x)) = (y - F(x))^2/2$

We want to minimize $J = \sum_{i} L(y_i, F(x_i))$ by adjusting

$$F(x_1), F(x_2), \dots, F(x_n).$$

Notice that $F(x_1), F(x_2), \dots, F(x_n)$ are just some numbers. We can treat $F(x_i)$ as parameters and take derivatives.

$$\frac{\partial \sum_{i} L(y_i, F(x_i))}{\partial F(x_i)} = \frac{\partial L(y_i, F(x_i))}{F(x_i)} = F(x_i) - y_i$$

So we can interpret residuals as negative gradients:

$$-\frac{\sum_{i} L(y_i, F(x_i))}{\partial F(x_i)} = y_i - F(x_i)$$

Jiayu Zhou

How is this related to gradient descent?

$$F(x_i) = F(x_i) + h(x_i)$$

$$F(x_i) = F(x_i) + y_i - F(x_i)$$

$$F(x_i) = F(x_i) - 1 \frac{\partial \sum_i L(y_i, F(x_i))}{\partial F(x_i)}$$

$$\theta_i = \theta_i - \rho \frac{\partial J}{\partial \theta_i}$$

Regression

Gradient Boosting for Regression

How is this related to gradient descent?

For regression with square loss,

residual = negative gradient

fit h to residual = fit h to negative gradient

update F based on residual = update F based on negative gradient

How is this related to gradient descent?

For regression with square loss,

residual = negative gradient

fit h to residual = fit h to negative gradient

update F based on residual = update F based on negative gradient

So we are actually updating our model using gradient descent!

How is this related to gradient descent?

For regression with square loss,

residual = negative gradient

fit h to residual = fit h to negative gradient

update F based on residual = update F based on negative gradient

So we are actually updating our model using **gradient descent**! It turns out that the concept of **gradients** is more general and useful than the concept of **residuals**. So from now on, lets stick with gradients. The reason will be explained later.

Let us summarize the algorithm we just derived using the concept of gradients. Negative gradient:

$$-g(x_i) = -\frac{\partial \sum_i L(y_i, F(x_i))}{\partial F(x_i)} = y_i - F(x_i)$$

- \bullet start with an initial model, say, $F(x) = \frac{\sum_{i=1}^n y_i}{n}$
- iterate until converge:
 - calculate negative gradients $-g(x_i)$
 - fit a regression model h to negative gradients $-g(x_i)$
 - $F = F + \rho h$, where $\rho = 1$

Let us summarize the algorithm we just derived using the concept of gradients. Negative gradient:

$$-g(x_i) = -\frac{\partial \sum_i L(y_i, F(x_i))}{\partial F(x_i)} = y_i - F(x_i)$$

- start with an initial model, say, $F(x) = \frac{\sum_{i=1}^{n} y_i}{n}$
- iterate until converge:
 - calculate negative gradients $-g(x_i)$
 - fit a regression model h to negative gradients $-g(x_i)$
 - $F = F + \rho h$, where $\rho = 1$

The benefit of formulating this algorithm using gradients is that it allows us to consider other loss functions and derive the corresponding algorithms in the same way.

Loss Functions for Regression Problem

Why do we need to consider other loss functions? Isn't square loss good enough?

Loss Functions for Regression Problem

Why do we need to consider other loss functions? Isn't square loss good enough?

Square loss is not robust to outliers

• Outliers are heavily punished because the error is squared.

y_i	0.5	1.2	2	5*
$F(x_i)$	0.6	1.4	1.5	1.7
$L = (y - F)^2/2$	0.005	0.02	0.125	5.445

Jiayu Zhou

Consequence?

Loss Functions for Regression Problem

Why do we need to consider other loss functions? Isn't square loss good enough?

Square loss is not robust to outliers

Outliers are heavily punished because the error is squared.

y_i	0.5	1.2	2	5*
$F(x_i)$	0.6	1.4	1.5	1.7
$L = (y - F)^2/2$	0.005	0.02	0.125	5.445

 Consequence? Pay too much attention to outliers. Try hard to incorporate outliers into the model. Degrade the overall performance.

Jiayu Zhou

Loss Functions for Regression Problem

Absolute loss (more robust to outliers)

Regression

$$L(y, F) = |y - F|$$

Loss Functions for Regression Problem

Absolute loss (more robust to outliers)

$$L(y, F) = |y - F|$$

• Huber loss (more robust to outliers)

$$L(y,F) = \begin{cases} \frac{1}{2}(y-F)^2 & |y-F| \le \delta \\ \delta(|y-F| - \delta/2) & |y-F| > \delta \end{cases}$$

Loss Functions for Regression Problem

Absolute loss (more robust to outliers)

$$L(y, F) = |y - F|$$

Huber loss (more robust to outliers)

$$L(y,F) = \begin{cases} \frac{1}{2}(y-F)^2 & |y-F| \le \delta\\ \delta(|y-F|-\delta/2) & |y-F| > \delta \end{cases}$$

y_i	0.5	1.2	2	5*
$F(x_i)$	0.6	1.4	1.5	1.7
Square loss	0.005	0.02	0.125	5.445
Absolute loss	0.1	0.2	0.5	3.3
Huber loss ($\delta = 0.5$)	0.005	0.02	0.125	1.525

Regression with Absolute Loss

Negative (sub)gradient

$$-g(x_i) = -\frac{\partial \sum_i L(y_i, F(x_i))}{\partial F(x_i)} = \operatorname{sign}(y_i - F(x_i))$$

- lacktriangle start with an initial model, say, $F(x) = \frac{\sum_{i=1}^n y_i}{n}$
- iterate until converge:
 - calculate negative gradients $-g(x_i)$
 - ullet fit a regression model h to negative gradients $-g(x_i)$
 - $F = F + \rho h$, where $\rho = 1$

Regression with Huber Loss

Negative (sub)gradient

$$\begin{split} -g(x_i) &= -\frac{\partial \sum_i L(y_i, F(x_i))}{\partial F(x_i)} \\ &= \begin{cases} y_i - F(x_i) & |y_i - F(x_i)| \leq \delta \\ \delta \mathrm{sign}(y_i - F(x_i)) & |y_i - F(x_i)| > \delta \end{cases} \end{split}$$

- lacksquare start with an initial model, say, $F(x) = \frac{\sum_{i=1}^n y_i}{n}$
- iterate until converge:
 - calculate negative gradients $-g(x_i)$
 - fit a regression model h to negative gradients $-g(x_i)$
 - $F = F + \rho h$, where $\rho = 1$

Regression with loss function L: general procedure

Given any (sub)differentiable loss function L

- start with an initial model, say, $F(x) = \frac{\sum_{i=1}^{n} y_i}{\pi}$
- iterate until converge:
 - calculate negative gradients $-g(x_i) = -\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}$
 - fit a regression model h to negative gradients $-q(x_i)$
 - $F = F + \rho h$, where $\rho = 1$

Regression with loss function L: general procedure

Given any (sub)differentiable loss function L

- start with an initial model, say, $F(x) = \frac{\sum_{i=1}^{n} y_i}{n}$
- iterate until converge:
 - calculate negative gradients $-g(x_i) = -\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}$
 - ullet fit a regression model h to negative gradients $-g(x_i)$
 - ullet F=F+
 ho h, where ho=1

In general,

negative gradients
$$\neq$$
 residuals

We should follow negative gradients rather than residuals. Why?

Negative Gradient vs Residual: An Example

Huber loss

$$L(y,F) = \begin{cases} \frac{1}{2}(y-F)^2 & |y-F| \le \delta \\ \delta(|y-F|-\delta/2) & |y-F| > \delta \end{cases}$$

Update by Negative Gradient:

$$h(x_i) = -g(x_i) = \begin{cases} y_i - F(x_i) & |y_i - F(x_i)| \le \delta \\ \delta \operatorname{sign}(y_i - F(x_i)) & |y_i - F(x_i)| > \delta \end{cases}$$

Update by Residual:

$$h(x_i) = y_i - F(x_i)$$

Difference: negative gradient pays less attention to outliers.

Summary of the Section

- ullet Fit an additive model $F=\sum_t
 ho_t h_t$ in a forward stage-wise manner.
- In each stage, introduce a new regression tree h to compensate the shortcomings of existing model.
- The "shortcomings" are identified by negative gradients.
- For any loss function, we can derive a gradient boosting algorithm.
- Absolute loss and Huber loss are more robust to outliers than square loss.

Things not covered

How to choose a proper learning rate for each gradient boosting algorithm. See [Friedman, 2001]

Classification

Problem

Recognize the given hand written capital letter.

- Multi-class classification
- 26 classes. A,B,C,...,Z



Data Set

- http://archive.ics.uci.edu/ml/datasets/Letter+Recognition
- 20000 data points, 16 features (how do we extract features in this case?)

Feature Extraction

Statistical moments and edge counts



1	horizontal position of box	9	mean y variance
2	vertical position of box	10	mean x y correlation
3	width of box	11	mean of x * x * y
4	height of box	12	mean of x * y * y
5	total number on pixels	13	mean edge count left to right
6	mean x of on pixels in box	14	correlation of x-ege with y
7	mean y of on pixels in box	15	mean edge count bottom to top
8	mean x variance	16	correlation of y-ege with x

Feature Vector= (2,1,3,1,1,8,6,6,6,6,5,9,1,7,5,10)

Label = G

Model

- 26 score functions (our models): $F_A, F_B, F_C, \dots, F_Z$.
- $F_A(x)$ assigns a score for class A
- scores are used to calculate probabilities

$$P_A(x) = \frac{e^{F_A(x)}}{\sum_{c=A}^{Z} e^{F_c(x)}}$$

$$P_B(x) = \frac{e^{F_B(x)}}{\sum_{c=A}^{Z} e^{F_c(x)}}$$

$$\cdots$$

$$P_Z(x) = \frac{e^{F_Z(x)}}{\sum_{c=A}^{Z} e^{F_c(x)}}$$

predicted label = class that has the highest probability

• turn the label y_i into a (true) probability distribution $Y_c(x_i)$ For example $y_5 = G$, then

$$Y_A(x_5) = 0, Y_B(x_5) = 0, \dots, Y_G(x_5) = 1, \dots, Y_Z(x_5) = 0$$

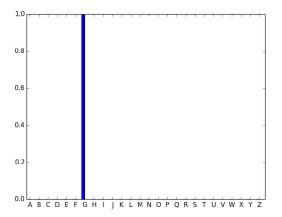


Figure: true probability distribution

30 / 58

• turn the label y_i into a (true) probability distribution $Y_c(x_i)$ For example $y_5 = G$, then

$$Y_A(x_5) = 0, Y_B(x_5) = 0, \dots, Y_G(x_5) = 1, \dots, Y_Z(x_5) = 0$$

② calculate the predicted probability distribution $P_c(x_i)$ based on the current model F_A, F_B, \ldots, F_Z .

$$P_A(x_5) = 0.03, P_B(x_5) = 0.05, \dots, P_G(x_5) = 0.3, \dots, P_Z(x_5) = 0.05$$

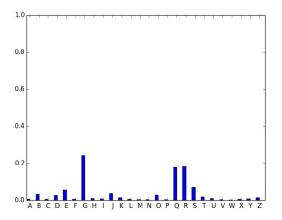


Figure: predicted probability distribution based on current model

• turn the label y_i into a (true) probability distribution $Y_c(x_i)$ For example $y_5=G$, then

$$Y_A(x_5) = 0, Y_B(x_5) = 0, \dots, Y_G(x_5) = 1, \dots, Y_Z(x_5) = 0$$

② calculate the predicted probability distribution $P_c(x_i)$ based on the current model F_A, F_B, \ldots, F_Z .

$$P_A(x_5) = 0.03, P_B(x_5) = 0.05, \dots, P_G(x_5) = 0.3, \dots, P_G(x_5) = 0.05$$

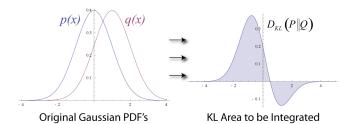
Solution and the predicted probability distribution. Here we use KL-divergence

Kullback-Leibler (KL) divergence

ullet Measures of the difference between two probability distributions P (true distribution) and Q (model distribution)

$$D_{\mathsf{KL}}(P||Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)}$$

$$= -\sum_{x} p(x) \log q(x) + \sum_{x} p(x) \log p(x) = H(P, Q) - H(P)$$



Kullback-Leibler Divergence

Example For a random variable $X=\{0,1\}$ assume two distributions P(x) and Q(x) with P(0)=1-r, P(1)=r and Q(0)=1-s, Q(1)=s:

$$D(P||Q) = (1-r)\log\frac{1-r}{1-s} + r\log\frac{r}{s}$$
$$D(Q||P) = (1-s)\log\frac{1-s}{1-r} + s\log\frac{s}{r}$$

If r = s

Kullback-Leibler Divergence

Example For a random variable $X=\{0,1\}$ assume two distributions P(x) and Q(x) with P(0)=1-r, P(1)=r and Q(0)=1-s, Q(1)=s:

$$D(P||Q) = (1-r)\log\frac{1-r}{1-s} + r\log\frac{r}{s}$$
$$D(Q||P) = (1-s)\log\frac{1-s}{1-r} + s\log\frac{s}{r}$$

If r = s then D(P||Q) = D(Q||P) = 0. If $r = \frac{1}{2}$ and $s = \frac{1}{4}$:

Kullback-Leibler Divergence

Example For a random variable $X = \{0,1\}$ assume two distributions P(x) and Q(x) with P(0) = 1 - r, P(1) = r and Q(0) = 1 - s, Q(1) = s:

$$D(P||Q) = (1-r)\log\frac{1-r}{1-s} + r\log\frac{r}{s}$$
$$D(Q||P) = (1-s)\log\frac{1-s}{1-r} + s\log\frac{s}{r}$$

If r=s then D(P||Q)=D(Q||P)=0. If $r=\frac{1}{2}$ and $s=\frac{1}{4}$:

$$D(P||Q) = \frac{1}{2}\log\frac{1/2}{3/4} + \frac{1}{2}\log\frac{1/2}{1/4} = 0.2075$$

$$D(Q||P) = \frac{3}{4}\log\frac{3/4}{1/2} + \frac{1}{4}\log\frac{1/4}{1/2} = 0.1887$$

Properties of the Kullback-Leibler Divergence

- $D(P||Q) \ge 0$
- D(P||Q) = 0 iff P(x) = Q(x) for all $x \in X$;
- Typically $D(P||Q) \neq D(Q||P)$

Multi-Class Classification via KL Divergence

Goal

- minimize the total loss (KL-divergence)
- for each data point, we wish the predicted probability distribution to match the true probability distribution as closely as possible

37 / 58

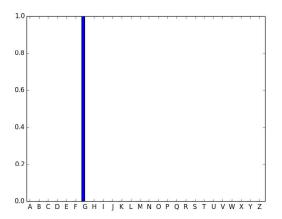


Figure: true probability distribution

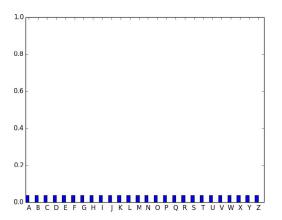


Figure: predicted probability distribution at round 0

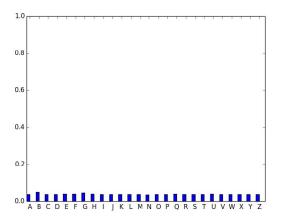


Figure: predicted probability distribution at round 1

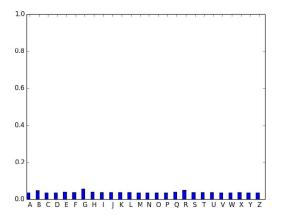


Figure: predicted probability distribution at round 2

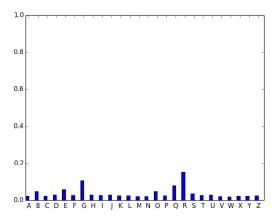


Figure: predicted probability distribution at round 10

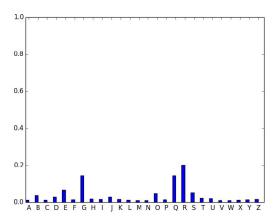


Figure: predicted probability distribution at round 20

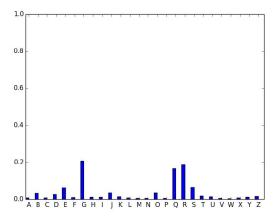


Figure: predicted probability distribution at round 30

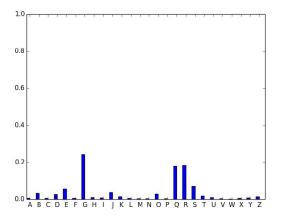


Figure: predicted probability distribution at round 40

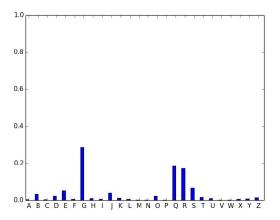


Figure: predicted probability distribution at round 50

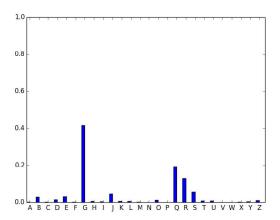


Figure: predicted probability distribution at round 100

The Classification Problem

Goal

- minimize the total loss (KL-divergence)
- for each data point, we wish the predicted probability distribution to match the true probability distribution as closely as possible
- we achieve this goal by adjusting our models F_A, F_B, \dots, F_Z .

Gradient Boosting for Regression: Review

Regression with loss function L: general procedure

Give any differentiable loss function Lstart with an initial model Fiterate until converge:

- calculate negative gradients $-g(x_i) = -\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}$
- fit a regression model h to negative gradients $-g(x_i)$
- \bullet $F = F + \rho h$

Difference between classification and regression

- \bullet F_A, F_B, \ldots, F_Z vs. F
- a matrix of parameters to optimize vs. a column of parameters to optimize

$F_A(x_1)$	$F_B(x_1)$	 $F_Z(x_1)$
$F_A(x_2)$	$F_B(x_2)$	 $F_Z(x_2)$
$F_A(x_d)$	$F_B(x_d)$	 $F_Z(x_d)$

• a matrix of gradients vs. a column of gradients

$\frac{\partial L}{\partial F_A(x_1)}$	$\frac{\partial L}{\partial F_B(x_1)}$	 $\frac{\partial L}{\partial F_Z(x_1)}$
$\frac{\partial L}{\partial F_A(x_2)}$	$\frac{\partial L}{\partial F_B(x_2)}$	 $\frac{\partial L}{\partial F_Z(x_2)}$
$\frac{\partial L}{\partial F_A(x_d)}$	$\frac{\partial L}{\partial F_B(x_d)}$	 $\frac{\partial L}{\partial F_Z(x_d)}$

start with an initial models $F_A, F_B, F_C, \dots, F_Z$ iterate until converge:

- ullet calculate negative gradients for class A: $-g_A(x_i) = -rac{\partial L(y_i, F(x_i))}{\partial F_A(x_i)}$
- ullet calculate negative gradients for class $B\colon -g_B(x_i) = -rac{\partial L(y_i,F(x_i))}{\partial F_B(x_i)}$
- ...
- ullet calculate negative gradients for class Z: $-g_Z(x_i) = -\frac{\partial L(y_i, F(x_i))}{\partial F_Z(x_i)}$
- ullet fit a regression model h_A to negative gradients $-g_A(x_i)$
- ullet fit a regression model h_B to negative gradients $-g_B(x_i)$
- ...
- ullet fit a regression model h_Z to negative gradients $-g_Z(x_i)$
- $\bullet \ F_A = F_A + \rho_A h_A$
- $\bullet \ F_B = F_B + \rho_B h_B$
- ...
- $F_Z = F_Z + \rho_Z h_Z$

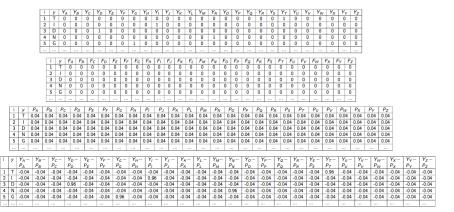
start with an initial models $F_A, F_B, F_C, \dots, F_Z$ iterate until converge:

- ullet calculate negative gradients for class $A\colon -g_A(x_i)=Y_A(x_i)-P_A(x_i)$
- ullet calculate negative gradients for class B:

$$-g_B(x_i) = Y_B(x_i) - P_B(x_i)$$

- ...
- calculate negative gradients for class Z: $-g_Z(x_i) = Y_Z(x_i) P_Z(x_i)$
- ullet fit a regression model h_A to negative gradients $-g_A(x_i)$
- ullet fit a regression model h_B to negative gradients $-g_B(x_i)$
- ...
- ullet fit a regression model h_Z to negative gradients $-g_Z(x_i)$
- $\bullet \ F_A = F_A + \rho_A h_A$
- $\bullet \ F_B = F_B + \rho_B h_B$
- ...
- $\bullet \ F_Z = F_Z + \rho_Z h_Z$

round 0



$$h_A(x) = \begin{cases} 0.98 & \text{feature 10 of } x \le 2.0 \\ -0.07 & \text{feature 10 of } x > 2.0 \end{cases}$$

$$h_B(x) = \begin{cases} -0.07 & \text{feature 15 of } x \le 8.0 \\ 0.22 & \text{feature 15 of } x > 8.0 \end{cases}$$

$$h_Z(x) = \begin{cases} -0.07 & \text{feature 8 of } x \leq 8.0\\ 0.82 & \text{feature 8 of } x > 8.0 \end{cases}$$

$$F_A = F_A + \rho_A h_A$$

$$F_B = F_B + \rho_B h_B$$

$$\cdots$$

$$F_Z = F_Z + \rho_Z h_Z$$

. . .

round 1

			i	у	YA		Yc	Y _D Y		YG		Y, Y			Y _M		-	Y _P Y		Y _S	YT		$Y_V \mid Y_V$			YZ					
			1	T	0	0	0	0 (0	0	0	0 1			0	0		0 (0	1		0 0		0	0					
			2	1	0	0	0	0 (0 (0	0		0 (0	0	0	0 (0	0		0 0	-	0	0					
			3	D	0	0	0	1 (0 0	0	0	0 1			0	0		0 (0	0		0 0		0	0					
			4	N	0	0	0	0 (0 0	0	0		0 (0	1	0	0 0		0	0	0	0 0	0	0	0					
			5	G	0	0	0	0 (0 0	1	0	0 1	0 (0	0	0	0	0 0	0	0	0	0	0 0	0	0	0					
																															_
i	y	FA	F _B	F	c	F_D	F _€	F_F	FG	F _H	F	F	,	F _K	FL	F_{M}	F_N	Fo	Fp	F	Q	F _R	F _S	F _T	Fu	F_V	F _W	F _X	Fy	Fz	
1	T	-0.08	-0.07	-0	.06	-0.07	-0.02	-0.02	-0.08	-0.03	2 -0.0	3 -0	03 -	0.06	-0.04	-0.08	-0.08	-0.0	7 -0.0	7 -0.	.02 -4	0.04 -	0.04 (0.59	-0.01	-0.07	-0.07	-0.05	-0.06	-0.07	f
2	1	-0.08				-0.07	-0.02		0.16	-0.0						-0.08	-0.08	-0.0	7 -0.0	7 -0.				0.07	-0.01	-0.07	-0.07	-0.05	-0.06	-0.07	
3	D	-0.08	0.23	-0	.06	-0.07	-0.02	-0.02	-0.08	-0.0	2 -0.0	3 -0	03 -	0.06	-0.04	-0.08	-0.08	-0.0	7 -0.0	7 -0.	.02 -1	0.04 -	0.04 -	0.07	-0.01	-0.07	-0.07	-0.05	-0.06	-0.07	
4	N	-0.08	-0.07	-0	.06	-0.07	-0.02	-0.02	0.16	-0.0	2 -0.0	3 -0	03 (0.26	-0.04	-0.08	0.3	-0.0	7 -0.0	7 -0.	.02 -4	0.04 -	0.04 -	0.07	-0.01	-0.07	-0.07	-0.05	-0.06	-0.07	7
5	G	-0.08	0.23	-0	.06	-0.07	-0.02	-0.02	0.16	-0.0	2 -0.0	3 -0	03 -	0.06	-0.04	-0.08	-0.08	-0.0	7 -0.0	7 -0.	.02 -1	0.04 -	0.04 -	0.07	-0.01	-0.07	-0.07	-0.05	-0.06	-0.07	ī
																															7
_																_															
	i	У	P _A	P _B	P_C	PD	PE	P_F	P_G	P_H	Pı	PJ	PK	PL	P _M	PN	Po	Pp	PQ	P_R	Ps	PT	Pu	P_V	Pw	PX	PY	Pz			
	1	TC	0.04 0	.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	1 0.04	0.07	0.04	0.04	0.04	0.04	0.04	0.04			
	2	1 0	0.04 0	.05	0.04	0.04	0.04	0.04	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	1 0.04	1 0.04	0.04	0.04	0.04	0.04	0.04	0.04			
	3	D C	0.04 0	.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	1 0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04			
	4	N C	0.04 0	.04	0.04	0.04	0.04	0.04	0.05	0.04	0.04	0.04	0.05	0.04	0.04	0.05	0.04	0.04	0.04	0.04	1 0.04	1 0.04	0.04	0.04	0.04	0.04	0.04	0.04			
	5	GC	0.04 0	.05	0.04	0.04	0.04	0.04	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	1 0.04	1 0.04	0.04	0.04	0.04	0.04	0.04	0.04			
																-		_		_			-								
i	y	YA-	Y _B -	Yc	-	Y _D -	Y _E -	Y _F -	Y _G -	Y _H -	- Y _I	- Y	j - [Y _K -	Y _L -	Y _M -	YN	- Y	- Y	-	$Y_Q -$	Y _R -	Ys -	Y _T -	Yu.	- Y _V	- Yu	v- Y	x - Y	y -	Yz -
		P_A	P_B	Pc	.	PD	P_E	P_F	P_G	PH	PI	P		PK	P_L	P _M	PN	Po	P	,	Po	P_R	Ps	PT	Pu	Pv			x P	Y	Pz
1	Т	-0.04	-0.04	-0.0	04	-0.04	-0.04	-0.04	-0.04	-0.0		04 -0	.04	0.04	-0.04	-0.04	-0.0	4 -0.	04 -0		-0.04	-0.04	-0.04	0.93	-0.0	4 -0.0	04 -0.	04 -0	.04 -0	.04	-0.04
2	T	-0.04	-0.05	-0.0	04	-0.04	-0.04	-0.04	-0.05	-0.0	4 0.9	5 -0	.04	0.04	-0.04	-0.04	-0.0	4 -0.	04 -0	.04	-0.04	-0.04	-0.04	-0.04	-0.0	4 -0.0	04 -0.	04 -0	.04 -0	.04 -	-0.04
3	D	-0.04	-0.05	-0.0	04	0.96	-0.04	-0.04	-0.04	-0.0	4 -0.0)4 -(.04 -	0.04	-0.04	-0.04	-0.0	4 -0.	04 -0	.04	-0.04	-0.04	-0.04	-0.04	-0.0	4 -0.0	04 -0.	04 -0	.04 -0	.04 -	-0.04
4	N	-0.04	-0.04	-0.0	04	-0.04	-0.04	-0.04	-0.05	-0.0	4 -0.0)4 -(.04	0.05	-0.04	-0.04	0.99	5 -0.	04 -0	.04	-0.04	-0.04	-0.04	-0.04	-0.0	4 -0.0	04 -0.	04 -0	.04 -0	.04 -	-0.04
5	G	-0.04	-0.05	-0.0	04	-0.04	-0.04	-0.04	0.95	-0.0	4 -0.0)4 -(.04	0.04	-0.04	-0.04	-0.0	4 -0.	04 -0	.04	-0.04	-0.04	-0.04	-0.04	-0.0	4 -0.0	04 -0.	04 -0	.04 -0	.04 -	-0.04
1																				\neg											
				1	_			1	1	1						1	1 /11			_			1	1	1						

$$h_A(x) = \begin{cases} 0.37 & \text{feature 10 of } x \leq 2.0 \\ -0.07 & \text{feature 10 of } x > 2.0 \end{cases}$$

$$h_B(x) = \begin{cases} -0.07 & \text{feature 14 of } x \leq 5.0 \\ 0.22 & \text{feature 14 of } x > 5.0 \end{cases}$$

$$h_Z(x) = \begin{cases} -0.07 & \text{feature 8 of } x \leq 8.0 \\ 0.35 & \text{feature 8 of } x > 8.0 \end{cases}$$

$$F_A = F_A + \rho_A h_A$$

$$F_B = F_B + \rho_B h_B$$

$$\cdots$$

$$F_Z = F_Z + \rho_Z h_Z$$

round 2

			1	T	0	0	0	0 0	0	0	0 0	0	0	0	0	0	0 0	0	0	0	1	0	0 0	0	0	0				
			2		0	0	0	0 0	0	0	0 1	. 0	0	0	0	0	0 0	0	0	0	0	0	0 0	0	0	0				
			3	D	0	0	0	1 0	0	0	0 0	0	0	0	0	0	0 0	0	0	0	0	0	0 0	0	0	0				
			4	N	0	0	0	0 0	0	0	0 0	0	0	0	0	1	0 0	0	0	0	0	0	0 0	0	0	0				
			5	G	0	0	0	0 0	0	1	0 0	0	0	0	0	0	0 0	0	0	0	0	0	0 0	0	0	0				
i	y	FA	F _B	Fc	F	D	FE	F _F	FG	F _H	F _I	F	F	K	FL	F _M	F _N	Fo	Fp	Fq	, ,	F _R	F _S	F _T	Fu	Fv	Fw	F _X	Fy	FZ
1	Т	-0.15	-0.14	-0.12		.14	-0.03	0.28	-0.14	-0.04	1.49	-0.0			0.08	-0.14	-0.17	-0.13	-0.13			0.11 -0	0.07 1	.05 (0.19	0.25	-0.16	-0.09	0.33	0.14
2	1	-0.15	0.16	-0.12	2 -0	.14	-0.03	-0.08	0.33	-0.04	-0.07	-0.0	7 -0.	.11 -	0.08	-0.14	-0.17	-0.13	-0.13	-0.0	14 -0	0.11 -0	0.07 -4	0.11 -	0.07	-0.15	-0.16	-0.09	-0.13 -	0.14
3	D	-0.15	0.16	-0.12	2 -0	.14	-0.03	-0.08	0.1	-0.04	-0.07	-0.0	7 -0.	.11 -	0.08	-0.14	-0.17	-0.13	-0.13	-0.0	14 0	.19 -0	0.07 -4	0.11 -	0.07	-0.15	-0.16	-0.09	-0.13 -	0.14
4	N	-0.15	-0.14	-0.12	2 -0	.14	-0.03	-0.08	0.1	-0.04	-0.07	-0.0	7 0.	46 -	0.08	-0.14	0.5	-0.13	-0.13	-0.0	14 -0	0.11 -0	0.07 -4	0.11 -	0.07	-0.15	0.25	-0.09	-0.13 -	0.14
5	G	-0.15	0.16	-0.12	2 -0	.14	-0.03	-0.08	0.33	-0.04	-0.07	-0.0	7 -0.	.11 -	0.08	-0.14	-0.17	-0.13	-0.13	-0.0	14 0	.19 -0	0.07 -4	0.11 -	0.07	-0.15	-0.16	-0.09	-0.13 -	0.14
													٠.																	
	i	y i	$P_A \mid P$			P_D	PE	P_F	P_G	P _H		P_J	PK	P_L	P _M	P _N	Po	Pp	PQ	P_R	P _S		PU	P_V	P_W	PX	Py	Pz		
	1	T 0	.03 0.0	0.	03	0.03	0.03	0.04	0.03	0.03	0.15	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.09	0.04	0.04	0.03	0.03	0.05	0.03		
	2	1 0	.04 0.1			0.04	0.04	0.04	0.06					0.04	0.04		0.04	0.04		0.04	0.04		0.04	0.04	0.04	0.04		0.04		
	3	D 0	.04 0.1	05 0.	04	0.04	0.04	0.04	0.05		0.04			0.04	0.04		0.04	0.04	0.04	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04		
	4	N 0	.03 0.1	0.	03	0.03	0.04	0.04	0.04		0.04			0.04	0.03		0.03	0.03	0.04	0.04	0.04		0.04	0.03	0.05	0.04	0.03	0.03		
	5	G 0	.03 0.1	0.0	04	0.04	0.04	0.04	0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.03	0.04	0.04	0.04	0.05	0.04	0.04	0.04	0.04	0.03	0.04	0.04	0.04		
i	y	YA-	Y _B -	Yc-	Y	-	Y _E -	Y _F -	Y _G -	Y _H -	Y1 -	Y	- Y	'κ-	Y _L -	Y _M -	Y _N -	Yo-	- Y _P	- Y	'q-	$Y_R -$	Y5 -	Y _T -	Yu-	- Y _V	- Y _W	- Y _X	- Y _Y -	- Yz -
		P_A	P _B	P_C	Po	,	P_E	P_F	P_G	PH	Pı	PJ	P	ĸ	P_L	P _M	PN	Po	Pp	P	9	P_R	P_S	PT	Pu	P_V	Pw	PX	PY	Pz
1	Т	-0.03	-0.03	-0.03	-0.	03	-0.03	-0.04	-0.03	-0.03	-0.15	-0.0	03 -0	0.03	-0.03	-0.03	-0.03	-0.03	3 -0.0)3 -(0.03	-0.03	-0.03	0.91	-0.04	4 -0.0	4 -0.0	3 -0.0	3 -0.05	-0.03
2	1	-0.04	-0.05	-0.04	-0.	04	-0.04	-0.04	-0.06	-0.04	0.96	-0.0)4 -(0.04	-0.04	-0.04	-0.04	-0.0	4 -0.0)4 -(0.04	-0.04	-0.04	-0.04	-0.04	4 -0.0	4 -0.0	4 -0.0	4 -0.04	4 -0.04
3	D	-0.04	-0.05	-0.04	0.9	96	-0.04	-0.04	-0.05	-0.04	-0.04	-0.0)4 -0	0.04	-0.04	-0.04	-0.04	-0.0	4 -0.0)4 -(0.04	-0.05	-0.04	-0.04	-0.04	4 -0.0	4 -0.0	4 -0.0	4 -0.04	4 -0.04
4	N	-0.03	-0.03	-0.03	-0.	03	-0.04	-0.04	-0.04	-0.04	-0.04	-0.0	04 -0	0.06	-0.04	-0.03	0.94	-0.0	3 -0.0	3 -0	0.04	-0.04	-0.04	-0.04	-0.04	4 -0.0	3 -0.0	5 -0.0	4 -0.03	3 -0.03
5	G	-0.03	-0.05	-0.04	-0.	04	-0.04	-0.04	0.94	-0.04	-0.04	-0.0	34 -0	0.04	-0.04	-0.04	-0.03	-0.0	4 -0.0)4 -(0.04	-0.05	-0.04	-0.04	-0.04	4 -0.0	4 -0.0	3 -0.0	4 -0.04	-0.04
_	_				_											_		-		_										

round 100

	i	y YA	Y _B	Yc	Y _D Y _I	YF	YG	Y _H Y	i Y	Yĸ	YL	Y _M	YN	Yo '	Yp	Yo	Y _R	Ys	Y _T	Yu Y	V Yv	v Y _X	Y _Y	Yz					
	1	T 0	0	0	0 0	0	0	0 (0 0	0	0	0	0	0	0	0	0	0	1	0	0 0	0	0	0					
	2	1 0	0	0	0 0	0	0	0 :	. 0	0	0	0	0	0	0	0	0	0	0	0	0 0	0	0	0					
	3	D 0	0	0	1 0	0	0	0 (0 0	0	0	0	0	0	0	0	0	0	0	0	0 0	0	0	0					
	4	N 0	0	0	0 0	0	0	0 (0	0	0	0	1	0	0	0	0	0	0	0	0 0	0	0	0					
	5	G 0	0	0	0 0	0	1	0 (0	0	0	0	0	0	0	0	0	0	0	0	0 0	0	0	0					
			1																				1						
																							_						
i y F _A	F _B	Fc	F _D	FE	F _F	F _G	F _H	F	F	F	۲ .	FL	F_{M}	F _N	F	0	Fp	FQ	F	R	F _S	F _T	Fu	F _V	F _W	F _X	F _Y	Fz	
1 T -3.26	-2.7	-2.2	-2.22	-2.48	-0.31	-2.77	-1.19	2.77	0.1	-1.4	19 -	1.02	-1.64	-0.8	-2	2.4	-3.57	-0.9	-2	.45 -	0.2 4	.61	0.5	-0.71	-1.21	-0.24	0.49	-1.6	56
2 I -1.64	-1.09	-2.29	-1.8	0.45	-0.43	2.14	-1.56	1.19	1.09	-1.	5 -	0.5	-3.64	-3.98	-0.	.39	-2.3	1.42	-0	.59 0	.27 -2	2.88 -	1.96	-1.67	-4.38	-2.0	-2.95	-1.7	76
3 D -2.45	0.18	-3.01	0.18	-2.79	-1.7	-2.21	0.43	-1.12	0.32	0.6	7 -2	2.16	-2.91	-2.76	-1.	.92	-3.04	-1.4	7 -0	.48 -1	L.48 -1	1.25 -	2.25	-3.23	-4.38	0.17	-2.95	-2.6	55
4 N -3.95	-3.38	-0.22	-0.94	-1.33	-1.38	-1.22	-0.12	-2.33	-3.13	0.5	8 -0	0.65	-0.25	2.96	-2.	.84	-1.82	0.19	0.	55 -1	1.22 -1	1.25	0.45	-1.8	0.11	-0.69	-1.6	-3.7	78
5 G -3.14	-0.04	-2.37	-0.78	0.02	-2.68	2.6	-1.48	-1.93	0.42	-1.4	14 -	1.45	-3.36	-3.98	-0.	.94	-3.42	1.84	1.	44 0	.62 -1	1.25 -	1.33	-4.41	-4.71	-2.62	2 -2.15	-1.0	9
	i y P _A P _B P _C P _D P _E P _E P _C P _H P _I P _I P _I P _I P _I P _I P _N P _I P _D P _T P _U P _V P _W P _X P _Y P _Z																												
						P_G	P _H	Pı		P _K	PL	P _M	P _N					P_R	Ps	P _T		P_V	P _W	PX	P _Y	PZ			
	0 0				0.01	0			0.01	0	0	0	0	0	0				0.01	0.79	0.01	0	0	0.01	0.01	0	ļ		
2 1 0.						0.32					0.02	0	0	0.03					0.05	0		0.01	0	0	0	0.01			
3 D 0.						0.01					0.01	0	0.01						0.02		0.01	0	0	0.11	0	0.01			
	0 0					0.01	0.03	0			0.02		0.59						0.01	0.01			0.03	0.02	0.01	0			
5 G	0.0	0 0	0.0	1 0.03	8 0	0.42	0.01	0	0.05	0.01	0.01	0	0	0.03	1 0	0.	19 (0.13	0.06	0.01	0.01	0	0	0	0	0.01	ļ		
																											J		
- I I V	V	V	V	V	V	V	V	I W	LV	Lv	_	V	V	T V	- 1	V	l v	1 1/	_	V	V	W	I W	I V	LV		v 1	,	V
			Y _D -	Y _E -	Y _F -	Y _G -						Y _L -	Y _M -			Y ₀ -	Yp-			Y _R -	Y ₅ -	YT-				w-		Y-	Yz -
			PD	PE	P _F	P _G	P _H	PI	PJ	PA		PL	P _M	PN		P _O	Pp	P		PR	Ps	PT	Pu	Pv				Y	Pz
			-0	-0	-0.01	-0.32	-0	-0.13				-0	-0	-0		0.03	-0	-0		-0	-0.01	0.21	-0.0		-0			0.01	-0
			-0.01	-0.06	-0.02		-0.01	0.88				-0.02	-0	-0		0.03	-0			-0.02	-0.05	-0	-0.0				0 -1		-0.01
			0.89	-0.01	-0.02	-0.01	-0.14	-0.03				-0.01	-0	-0.0		0.01	-0			-0.05	-0.02	-0.03					0.11 -		-0.01
			-0.01	-0.01	-0.01	-0.01	-0.03	-0	-0	-0.		-0.02	-0.02			0	-0			-0.05	-0.01	-0.01	-0.0					0.01	-0
	-0.03	-0	-0.01	-0.03	-0	0.58	-0.01	-0	-0.0	5 -0.	.01	-0.01	-0	-0	-	-0.01	-0	-0	.19	-0.13	-0.06	-0.01	-0.0	1 -0	-0		-0 -1	,	-0.01
						***		***	***																				***