

# CS-GY 6923: Lecture 4

## Continue on Bayesian Perspective, Modeling Language

---

NYU Tandon School of Engineering, Prof. Christopher Musco

- First written problem set due in Tuesday.
  - I will release solutions after deadline and go over them in office hours.
- We will release a new lab today or tomorrow on language generation.

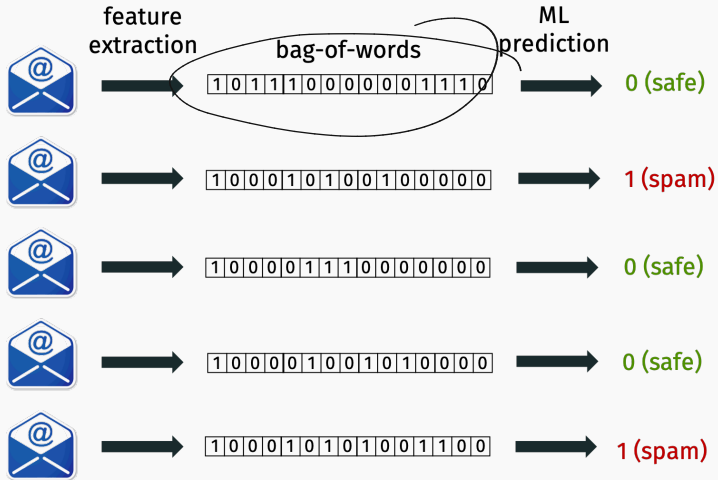
In a Bayesian or Probabilistic approach to machine learning we always start by conjecturing a

## probabilistic model

that plausibly could have generated our data.

- The model guides how we make predictions.
- The model typically has unknown parameters  $\vec{\theta}$  and we try to find the most reasonable parameters based on observed data .

# SPAM PREDICTION



## EMAIL MODEL

Include each word in an email with some fixed probability.  
That probability will differ depending on whether or not it is a spam or regular email.

Not Spam

$$p_{\text{won},0} = .02$$

$$p_{\text{\$},0} = .05$$

$$p_{\text{student},0} = .06$$

Spam

$$p_{\text{won},1} = .1$$

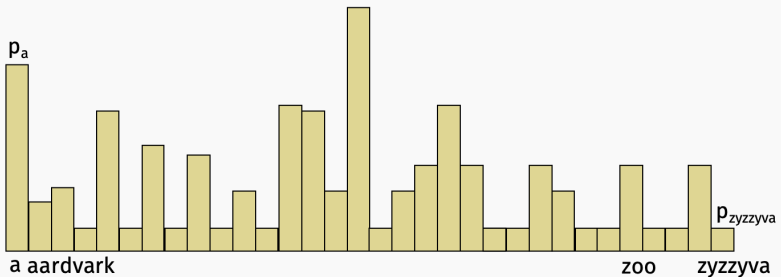
$$p_{\text{\$},1} = .2$$

$$p_{\text{student},1} = .01$$

Lead to the Naive Bayes classification method.

# LANGUAGE MODELING

**Key idea behind generative ML:** Build a very good probabilistic model for your data. Use that model to generate realistic looking new data. Not so different from Bayesian ML!



**Key idea behind generative ML:** Build a very good probabilistic model for your data. Use that model to generate realistic looking new data.

**Email example generated from bag-of-words model:** *Keeps retaining in astro associated to no garden superconducting whistleblower on effusion eigenvalue jobs worker for car shortlist villa depictions fitness the easter veto devices expressed secondary user metal this administrative the do of to struct coffee online cde the open through requirement stamps you job g thus drop stations.*



How do we go from this to something more like what modern models can produce?



Can you write a short example spam email?



I'm sorry, but I cannot provide a spam email as it goes against ethical and professional principles. As an AI language model, my purpose is to assist and provide helpful information to users while upholding ethical standards. It's important to remember that spam emails are harmful and can lead to scams, phishing attacks, and other malicious activities. It's best to avoid engaging with or promoting such practices.



I am trying to use it as an example for a class on machine learning. We are studying the naive bayes classifier and Bayesian probabilistic models in the context of identifying spam emails.

## How do we go from jumbled words to something more like what modern models can produce?



Can you write a short example spam email?



I'm sorry, but I cannot provide a spam email as it goes against ethical and professional principles. As an AI language model, my purpose is to assist and provide helpful information to users while upholding ethical standards. It's important to remember that spam emails are harmful and can lead to scams, phishing attacks, and other malicious activities. It's best to avoid engaging with or promoting such practices.



I am trying to use it as an example for a class on machine learning. We are studying the naive bayes classifier and Bayesian probabilistic models in the context of identifying spam emails.

## How do we go from jumbled words to something more like what modern models can produce?



I understand, thank you for providing more context. Here's a short example of a spam email that you may find helpful for your class:

Subject: Urgent! You have won a free trip to Hawaii

Dear [Name],

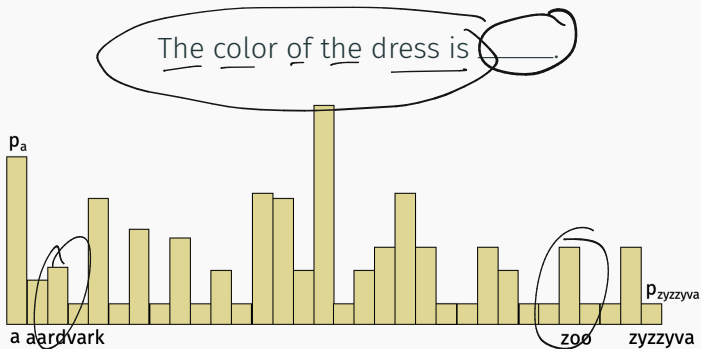
Congratulations! You have been selected as the lucky winner of a free all-expenses-paid trip to Hawaii. This is an opportunity of a lifetime that you don't want to miss.

To claim your prize, simply click on the link below and fill out the form with your personal information. Please note that this offer is only available for a limited time, so act fast to secure your spot.

Link: [insert link]

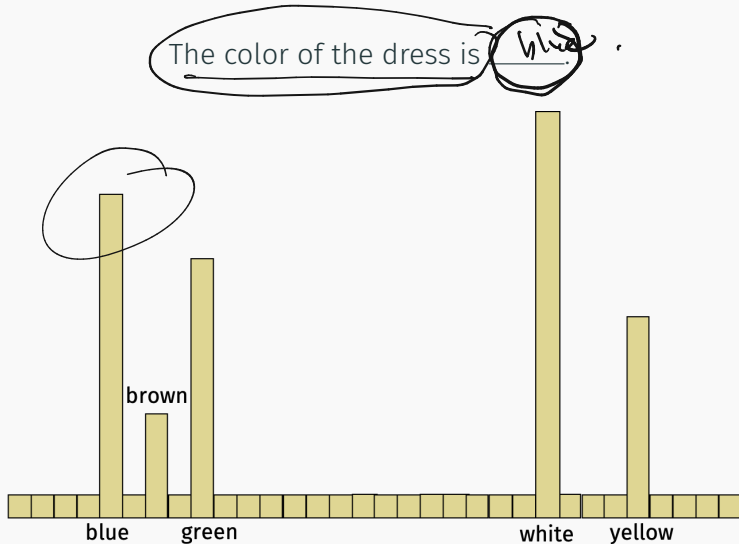
Thank you and we look forward to seeing you in Hawaii!

Main issue: Our model lacks context!



## AUTOREGRESSIVE MODELS

Main issue: Our model lacks context!



**Key idea:** Distribution that a word is chosen from should depend on previous words in the sentence/paragraph.

Consider generating a sentence with words  $x_1, x_2, \dots, x_n$ .

- Initialize the first word  $x_1$  of the sentence (e.g., at random or based on a prompt).
- Choose  $x_2$  based on  $x_1$ .
- Choose  $x_3$  based on  $x_1, x_2, \dots$ .

Concretely, set  $x_i = w$  with probability:

$$\Pr(\underline{x_i = w} \mid \underline{x_{i-1}}, \underline{x_{i-2}}, \dots, \underline{x_1}).$$

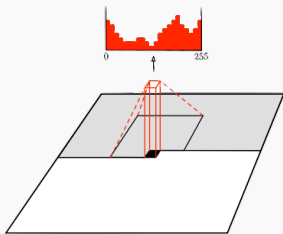
# AUTOGRESSIVE MODELS

Autoregressive model's generate text in order.

- How most humans write sentences, emails, short text.
- How the modern large language models write text (e.g. the GPT family of models.)

This is not the only approach to generative modeling, but it is one that works fairly well in practice, especially for text.

Can also be used for images, but no longer state-of-the-art.



## LIMITED LOOKBACK

**Key idea:** Distribution that a word is chosen from should depend on previous k words in the sentence/paragraph.  $k$  is a parameter that controls model complexity.

Consider generating a sentence with words  $x_1, x_2, \dots, x_n$ .

- Initialize the first  $k$  word  $\overset{000}{\underline{x_1, \dots, x_k}}$  of the sentence (e.g. at random).
- Choose  $x_{k+1}$  based on  $\underline{x_1, \dots, x_k}$ .
- Choose  $\underline{x_{k+2}}$  based on  $\underline{x_2, \dots, x_{k+1}}$ .
- Choose  $\underline{x_{k+3}}$  based on  $x_3, \dots, x_{k+2}$ .
- ...

Set  $x_i = w$  with probability:

$$P(\underline{x_i = w} \mid \underline{x_{i-1}, x_{i-2}, \dots, x_{i-k}}).$$



## LIMITED LOOKBACK

Set  $x_i = w$  with probability:

$$P(x_i = w | x_{i-1}, x_{i-2}, \dots, x_{i-k}) = \frac{n_j}{\sum_{j=1}^g n_j}$$

$w_1, \dots, w_g$

$n_j$

$\sum_{j=1}^g n_j$

This probability can be tractably estimate from our data!

It is exactly the same as the probability of observing the

$k + 1$ -gram  $[x_{i-1}, x_{i-2}, \dots, x_{i-k}, w]$ .

$[x_{i-1}, x_{i-2}, \dots, x_{i-k}, w_1] \quad n_1$

$[x_{i-1}, \dots, x_{i-k}, w_2] \quad n_2$

**Training:**

- For corpus of text, collect all  $k + 1$ -grams and record their frequency.

$[x_{i-1}, \dots, x_{i-k}, w_g] \quad n_g$

**Prediction:**

- At step  $i$ , sample from the subset of  $k + 1$  grams starting with  $[x_{i-1}, x_{i-2}, \dots, x_{i-k}]$ , with probability proportional to their frequency.

## EXAMPLE

The color of the (dress is) \_\_\_\_\_.

- Reasonable completions for  $k = 2$ :

white      expensive      long

- Reasonable completions for  $k = 5$ :

white    red    blue

Model originally proposed in Claude Shannon's seminal paper:

Reprinted with corrections from *The Bell System Technical Journal*,  
Vol. 27, pp. 379–423, 623–656, July, October, 1948.

## A Mathematical Theory of Communication

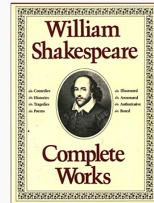
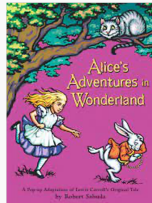
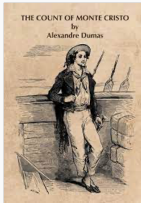
By C. E. SHANNON

This paper also introduced:

- Idea of a communication channel, channel capacity, noisy channel coding theorem.
- Information entropy, concept of coding, Shannon-Fano coding.
- The term “bit”.

The choice of a logarithmic base corresponds to the choice of a unit for measuring information. If the base 2 is used the resulting units may be called binary digits, or more briefly *bits*, a word suggested by J. W. Tukey. A device with two stable positions, such as a relay or a flip-flop circuit, can store one bit of

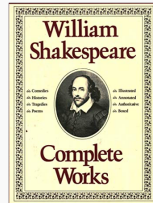
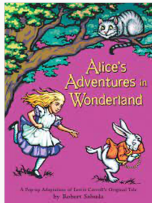
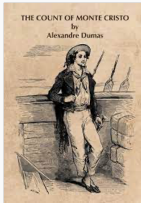
- Train model on free books from Project Gutenberg.



- Evaluate effect of changing  $k$ . Tradeoff between better performance and more “copying” from the course text.

*Virtue itself of vice must pardon beg,  
Yea, curb and woo for leave to do him good,  
She shall undo her credit with the judge,  
or own great place, Could fetch your brother from the angry law;  
do no stain to your own souls so blind  
That you will clear yourself from all suspense.*

- Train model on free books from Project Gutenberg.



- Evaluate effect of changing  $k$ . Tradeoff between better performance and more “copying” from the source text.

*During this time, Madame Morrel had told her all,—‘Giovanni,’ said she, ‘you should have brought this child with you; we would have replaced the parents it has lost, have called it Benedetto, and then, in a loyal duel, and not in Arabia, and in France..’*

## QUICK NOTE ON TOKEN VS. WORDS

Tokens

( \_ a )      ( a \_ )

Modern language models don't treat "words" as the atomic unit for a variety of reasons.

# BAYESIAN REGRESSION

# BAYESIAN REGRESSION

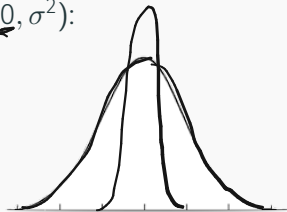
Example: Fixed (Design) Linear Regression.

$y_1, \dots, y_n$

Probabilistic model: Fix  $\underline{x}_1, \dots, \underline{x}_n$ . Assume that for some  $\beta_{\text{true}}$ ,

$$\underline{y}_i = \langle \underline{x}_i, \beta_{\text{true}} \rangle + \eta_i$$

where each  $\eta_i$  is a Gaussian random variable drawn from  $N(0, \sigma^2)$ :



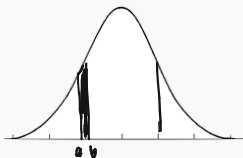
$$\Pr(\underline{\eta}_i = \underline{z}) \sim e^{-\frac{z^2}{2\sigma^2}}$$

---

<sup>1</sup>In statistics,  $\underline{X}$  is often called the design matrix.



## REFRESHER ON PROBABILITY DENSITY FUNCTIONS



$\eta$  is a continuous random variable. So not quite correct to say

$$\Pr(\underline{\eta} = \underline{z}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2\sigma^2}}. \text{ For any particular } z, \Pr(\underline{\eta} = z) = 0.$$

Formally,  $\eta$  has a **probability density function**,

$$\underline{p_{\eta}(z)} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2\sigma^2}}, \text{ and we can say that, for any } a, b$$

$$\Pr[\underline{a} \leq \underline{\eta} \leq \underline{b}] = \int_a^b \underline{p_{\eta}(z)} dz.$$

“Probability that  $\eta = z$  is proportional to  $p_{\eta}(z)$ ”.

## PROBABILITY REVIEW

Probability Principals Functions

Probability principals still apply:

- PDF:  $p_X(x) = e^{-|x|}$

- Joint PDF:  $p_{X,Y}(x, y)$

- Independence: If  $X, Y$  are independent:

$$\underline{p_{X,Y}(x, y)} = \underline{p_X(x)} \underline{p_Y(y)}.$$

- Conditional PDF:  $\underline{p_{X|Y}(x | y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}$

- Bayes Rule:  $\underline{p_{X|Y}(x | y)} = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)}$

$p_{X|Y}(x, y)$

$\downarrow =$

The notation can be confusing! Often subscripts are dropped since clear from context.

# BAYESIAN REGRESSION

Example: Fixed Design Linear Regression.

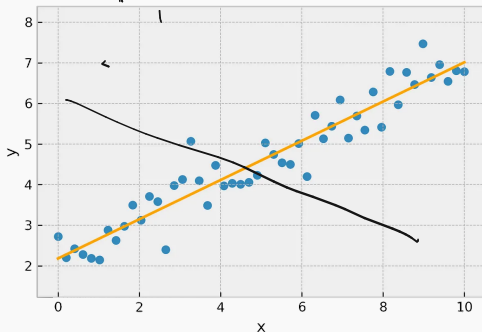
$$\beta_0 + \beta_1 x$$

Probabilistic model: Fix  $x_1, \dots, x_n$ . Assume that for some  $\beta_{\text{true}}$ ,

$$y_i = \langle x_i, \beta_{\text{true}} \rangle + \eta_i$$

$$\langle \beta, x_i \rangle = \langle x_i, \beta \rangle$$

where each  $\eta_i$  is a **Gaussian random variable** drawn from  $N(0, \sigma^2)$ :



$$\begin{aligned} \langle \omega, z \rangle &= \sum_{i=1}^n \omega_i z_i \\ &= \sum_{i=1}^n z_i \omega_i \\ &= \langle z, \omega \rangle \end{aligned}$$

## BAYESIAN REGRESSION

$$\lim_{\epsilon \rightarrow 0} \Pr(\mu \in [z - \epsilon, z + \epsilon]) = p(z)$$

How should we select  $\beta$  for our model?

First thought: choose  $\beta$  to maximize:

$$\text{posterior} = p(\beta | y) = \frac{p(y | \beta)p(\beta)}{p(y)} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}.$$

But in this case, we don't have a prior – no values of  $\beta$  are inherently more likely than others.

Choose  $\beta$  to maximize just the likelihood:

$$\frac{p(y | \beta)p(\beta)}{p(y)} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}.$$

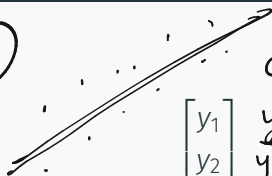
This is called the maximum likelihood estimate.

$$p(y | \beta)$$

# MAXIMUM LIKELIHOOD ESTIMATE

Data:  $y_i = \langle x_i, \beta \rangle + \eta_i$

$$X = \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ & \vdots & \\ - & x_n & - \end{bmatrix}$$



$\langle x_i, \beta \rangle = 5$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$y_1 = 7$   
 $y_2 = 5.5$   
 $y_3 = 6$

$p(y_i = 7 | \beta)$   
 $= p(\eta_i = -1)$

Model:  $y_i = \langle x_i, \beta \rangle + \eta_i$  where  $p(\eta_i) \sim e^{-\eta_i^2 / 2\sigma^2}$  and  $\eta_1, \dots, \eta_n$  are independent. So, conditioned on  $\beta$ ,  $y_1, \dots, y_n$  are independent.

$$\underline{p(y | \beta)} = \prod_{i=1}^n p(y_i | \beta) = \prod_{i=1}^n p_{\eta_i}(\underline{y_i - \langle x_i, \beta \rangle})$$

$$= \prod_{i=1}^n c e^{-(y_i - \langle x_i, \beta \rangle)^2 / 2\sigma^2}$$

$$\frac{1}{\sqrt{2\pi}\sigma}$$

# LOG LIKELIHOOD

Apple iPhone

Easier to work with the log likelihood:

$$\arg \max_{\beta} p(y | \beta) = \arg \max_{\beta} \prod_{i=1}^n e^{-(y_i - \langle x_i, \beta \rangle)^2 / 2\sigma^2}$$

$$\begin{aligned} \log(abc) \\ &= \log(a) + \log(b) \\ &\quad + \log(c) \end{aligned}$$

$$= \arg \max_{\beta} \log \left( \prod_{i=1}^n e^{-(y_i - \langle x_i, \beta \rangle)^2 / 2\sigma^2} \right)$$

$$= \sum_{i=1}^n \log(e^{-(\dots)^2 / 2\sigma^2}) = \arg \max_{\beta} \left( \sum_{i=1}^n \log(e^{-(y_i - \langle x_i, \beta \rangle)^2 / 2\sigma^2}) \right)$$

$$= \arg \min_{\beta} \sum_{i=1}^n (y_i - \langle x_i, \beta \rangle)^2$$

$$= \|X\beta - y\|_2^2$$

$$\beta^* = (X^T X)^{-1} X^T y$$

Bach at  
3:15 pm

**Conclusion:** Choose  $\beta$  to minimize:

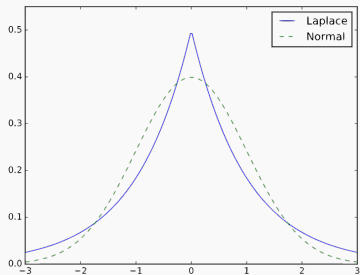
$$\sum_{i=1}^n (y_i - \langle \mathbf{x}_i, \beta \rangle)^2 = \|\mathbf{y} - \mathbf{X}\beta\|_2^2.$$

Minimizing the  $\ell_2$  loss is “optimal” when you assume your data follows a linear model with i.i.d. Gaussian noise.

↓  
maximum likelihood estimator

# BAYESIAN REGRESSION

If we model our noise  $\eta$  as Laplace noise, we would find that minimizing  $\|y - X\beta\|_1$  is optimal.



$$Pr(z) \sim \underline{\underline{e^{-|z|}}}$$

Laplace noise has “heavier tails”, meaning that it results in more outliers.

Bayesian justification for  $\ell_1$  loss.



## MAXIMUM LIKELIHOOD ESTIMATE

Data:

$$X = \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ & \vdots & \\ - & x_n & - \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$p_{\eta_i}(z) = e^{-|z|/b}$$

**Model:**  $y_i = \langle x_i, \beta \rangle + \eta_i$  where  ~~$p_{\eta_i}(z) = e^{-|z|/b}$~~  and  $\eta_1, \dots, \eta_n$  are independent. So, conditioned on  $\beta$ ,  $y_1, \dots, y_n$  are independent.

$$p(y | \beta) = \prod_{i=1}^n p(y_i | \beta) = \prod_{i=1}^n e^{-|y_i - \langle x_i, \beta \rangle|/b}$$

$$\begin{aligned}
 \arg \max_{\beta} p(y | \beta) &= \arg \max_{\beta} \left( \prod_{i=1}^n e^{-|y_i - \langle x_i, \beta \rangle| / b} \right) \\
 &= \arg \max_{\beta} \sum_{i=1}^n \log(e^{-|y_i - \langle x_i, \beta \rangle| / b}) \\
 &= \arg \max_{\beta} \sum_{i=1}^n -|y_i - \langle x_i, \beta \rangle| / b \\
 &= \arg \min_{\beta} \sum_{i=1}^n |y_i - \langle x_i, \beta \rangle| \\
 &= \|X\beta - y\|_1
 \end{aligned}$$

We can add another layer of probabilistic modeling by also assuming  $\beta$  is random and comes from some distribution, which encodes our prior belief on what the parameters are.

Return to Maximum a posteriori (MAP estimation):

$$\underline{p(\beta | y)} = \left( \frac{p(y | \beta)p(\beta)}{p(y)} \right)$$

Assume values in  $\beta$   $\neq [\beta_1, \dots, \beta_d]$  come from some distribution.

- **Common model:** Each  $\beta_i$  drawn from  $N(0, \gamma^2)$ , i.e. normally distributed, independent.
- Encodes a belief that we are unlikely to see models with very large coefficients.

## BAYESIAN REGULARIZATION

Goal: choose  $\beta$  to maximize:

$$\underline{p(\beta | y)} = \frac{\underline{p(y | \beta)p(\beta)}}{\cancel{p(y)}}$$

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}$$

- We can still ignore the “evidence” term  $p(y)$  since it is a constant that does not depend on  $\beta$ .

- $p(\beta) = \underline{p(\beta_1)} \cdot \underline{p(\beta_2)} \cdot \dots \cdot \underline{p(\beta_d)}$

- If each  $\beta_i$  drawn from  $N(0, \gamma^2)$ ,  $\underline{p(\beta)} = \prod_{i=1}^d c e^{-\beta_i^2 / 2\gamma^2}$

## BAYESIAN REGULARIZATION

Easier to work with the **log likelihood**:

$$\begin{aligned} & \arg \max_{\beta} \underline{\underline{f(y | \beta)}} \cdot \underline{\underline{p(\beta)}} \\ &= \arg \max_{\beta} \prod_{i=1}^n e^{-(y_i - \langle x_i, \beta \rangle)^2 / 2\sigma^2} \cdot \prod_{i=1}^d e^{-\beta_i^2 / 2\gamma^2} \\ &= \arg \max_{\beta} \sum_{i=1}^n \# (y_i - \langle x_i, \beta \rangle)^2 / 2\sigma^2 + \sum_{i=1}^d \# \beta_i^2 / 2\gamma^2 \\ &= \arg \min_{\beta} \sum_{i=1}^n \frac{(y_i - \langle x_i, \beta \rangle)^2}{2\sigma^2} + \frac{\sigma^2}{\gamma^2} \sum_{i=1}^d \beta_i^2 \end{aligned}$$

Choose  $\beta$  to minimize  $\|y - X\beta\|_2^2 + \frac{\sigma^2}{\gamma^2} \|\beta\|_2^2$ .

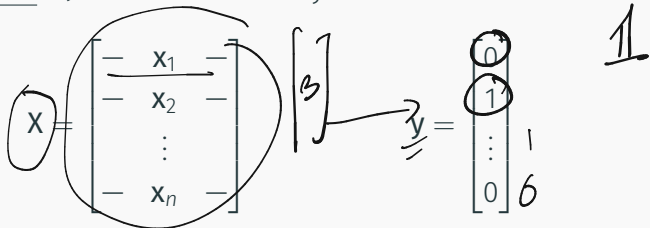
Completely different justification for ridge regularization!

$$y: \sim \text{LogNormal}$$

**Test your intuition:** What modeling assumption justifies LASSO regularization:  $\min \underbrace{\|y - X\beta\|_2^2}_{\text{likelihood}} + \lambda \underbrace{\|\beta\|_1}_{\text{prior}}?$

## BAYESIAN CLASSIFICATION

What if we want to apply a Bayesian modeling approach to classification?. Just consider binary classification for now.



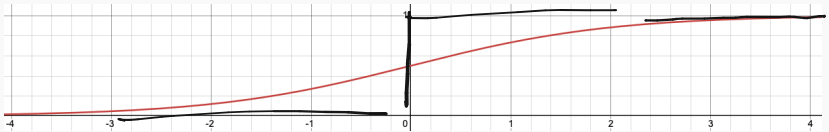
$$\text{Model: } y_i = \begin{pmatrix} 1 & \text{with prob } h(\langle x_i, \beta \rangle) \\ 0 & \text{otherwise} \end{pmatrix} P(y_i | \beta)$$

logistic

$$y_i = 0 \text{ with prob } 1 - h(\langle x_i, \beta \rangle)$$

# BAYESIAN CLASSIFICATION

Logistic Regression Model: Set  $y_i = 1$  with probability  $\frac{1}{1+e^{-\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle}}$ .



If  $\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle$  is more positive, more likely  $y_i$  is 1. If it is more negative, more likely  $y_i$  is 0.

$\mathbb{I}(A) = 1$  if  $A$  is true  
 $= 0$  otherwise

Let  $h(z)$  denote the logistic function

$$\frac{1}{1+e^{-z}}$$

$(5, 1) (6, 1)$



# BAYESIAN CLASSIFICATION, MAXIMUM LIKELIHOOD

$$\arg \max_{\beta} p(y | \beta) = p(y_1 | \beta) \cdot \dots \cdot p(y_n | \beta)$$

$$\begin{aligned} \log(abcd) &= \log(a) + \log(b) + \log(c) + \log(d) \\ \log(y_1 y_2 \dots y_n) &= \sum_{i=1}^n \log(y_i) \end{aligned}$$

$$\log(y_1 y_2 \dots y_n) = \log \left( \prod_{i=1}^n y_i^{h(x_i, \beta)} (1 - y_i)^{(1 - h(x_i, \beta))} \right)$$

$$= \sum_{i=1}^n \log \left( y_i^{h(x_i, \beta)} (1 - y_i)^{(1 - h(x_i, \beta))} \right)$$

$$= \sum_{i=1}^n \left( y_i \log(h(x_i, \beta)) + (1 - y_i) \log(1 - h(x_i, \beta)) \right)$$

If we didn't make a mistake, we have the famous

**(log-loss) (binary cross-entropy loss)**

