CS-UY 4563: Lecture 7
The Bayesian Perspective cont., Linear Classifiers

NYU Tandon School of Engineering, Prof. Christopher Musco

PROBABILISTIC MODELING

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 (more on this later in lecture).

PROBABILISTIC MODELING

Exercise: With a partner, come up with a probabilistic model for any one of the following data sets $(x_1, y_1), \ldots, (x_n, y_n)$.

1. For n **people**: each $x_i \in \{0, 1\}$ with zero indicating male, one indicating female. Each y_i is the height of the person in inches.
2. For n **NYC apartments**: each x_i is the size of the apartment in square feet. Each y_i is the monthly rent in dollars.
3. For n **students**: each $x_i \in \{Fresh., Soph., Jun., Sen.\}$ indicating class year. Each $y_1 \in \{0, 1\}$ with zero indicating the student has not taken machine learning, one indicating they have.

Record any unknown parameters of your model. What would be a guess for their values? How would you confirm or refine this guess using data?

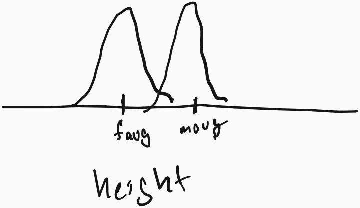
PROBABILISTIC MODELING

Dataset: $(x_1, y_1), \ldots, (x_n, y_n)$

$$\begin{array}{c} x \\ y \end{array} \begin{bmatrix} 0, & 1, & 1, & 0, & 0, & 1, & i \\ 60'' & 62'' & 70'' & \cdots & , & \cdot & \end{bmatrix}$$

Description: For n **people**: each $x_i \in \{0, 1\}$ with zero indicating male, one indicating female. Each y_i is the height of the person in inches. Parameters:

Model: σ^2, favg, mavg



(X, y)

$\downarrow$

0, 1 with probabilty 1/2, 1/2

$$y = \begin{cases} \text{If } x = 0, & y = \text{mavg} + N(0, \sigma^2) \\ \text{If } x = 1, & y = \text{favg} + N(0, \sigma^2) \end{cases}$$

using:

numpy.random.randn

PROBABILISTIC MODELING

Dataset: $(x_1, y_1), \ldots, (x_n, y_n)$

Description: For n **NYC apartments**: each x_i is the size of the apartment in square feet. Each y_i is the monthly rent in dollars.

Model:

min = 350 sq ft.

max = 5000 sq ft

$$y = c \cdot x + \mathcal{N}(0, \sigma^2)$$

$$\boxed{x \sim \text{Unif}[350, 5000]}$$

draw from uniform distribution

$(x, \_\_\_)$

$$\boxed{y = c \cdot x + \text{Unif}[-v, v]}$$

$$y = (c + \text{Unif}[-1, 1]) \cdot x$$

Using data: find $c = \$5$ $v = \$2$

PROBABILISTIC MODELING

Dataset: $(x_1, y_1), \ldots, (x_n, y_n)$

Description: For n **students**: each $x_i \in \{Fresh., Soph., Jun., Sen.\}$ indicating class year. Each $y_i \in \{0, 1\}$ with zero indicating the student has not taken machine learning, one indicating they have.

Model:

$$x \sim \text{unif}[\overset{1/4}{\text{fresh}}, \overset{1/4}{\text{soph}}, \overset{1/4}{\text{jun}}, \overset{1/4}{\text{sen.}}]$$

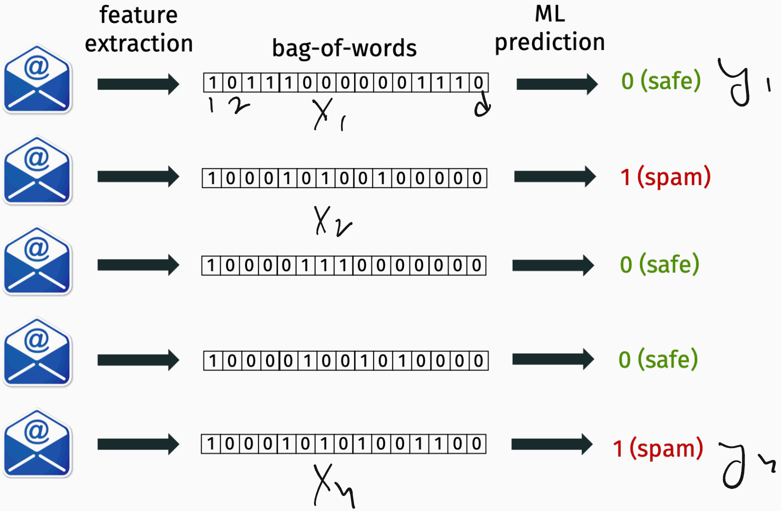
$$\begin{cases} \text{if } x = \text{fresh}, & y = 1 \text{ with prob. } 0 \\ \text{if } x = \text{soph}, & y = 1 \text{ with prob. } .05 \\ \ldots & \\ \text{if } x = \text{sen}, & y = 1 \text{ with prob. } .5 \end{cases}$$

NAIVE BAYES CLASSIFIER

Goal:

- Build a probabilistic model for a binary classification problem.
- Estimate parameters of the model.
- From the model derive a classification rule for future predictions (the Naive Bayes Classifier).

SPAM PREDICTION



feature extraction

bag-of-words

ML prediction

1	0	1	1	1	0	0	0	0	0	0	1	1	1	0

1 2 x_1 d

0 (safe) y_1

1	0	0	0	1	0	1	0	0	1	0	0	0	0	0

x_2

1 (spam)

1	0	0	0	0	1	1	1	0	0	0	0	0	0	0

0 (safe)

1	0	0	0	0	1	0	0	1	0	1	0	0	0	0

0 (safe)

1	0	0	0	1	0	1	0	1	0	0	1	1	0	0

x_n

1 (spam) y_n

Both target labels and data vectors are binary.

Probabilistic model for (bag-of-words, label) pair $(\mathbf{x}, y)$:

- Set $y = 0$ with probability $p(y = 0)$, $y = 1$ with probability $p(y = 1) = 1 - p(y = 0)$.
  - $p(y = 0)$ is probability an email is not spam (e.g. 99%).
  - $p(y = 1)$ is probability an email is spam (e.g. 1%).
- If $y = 0$, for each i, set $x_i = 1$ with prob. $p(x_i = 1 \mid y = 0)$.
- If $y = 1$, for each i, set $x_i = 1$ with prob. $p(x_i = 1 \mid y = 1)$.

$[0\ 1\ 1\ 1\ 0\ 1\ 0\ 1\ 1\ 1\ 1\ 1\ 1\ 1]$

$1 \quad \dots \quad m$

Unknown model parameters:

- $p(y = 0), p(y = 1)$,
- $p(x_1 = 1 \mid y = 0), \ldots, p(x_d = 1 \mid y = 0)$.
- $p(x_1 = 1 \mid y = 1), \ldots, p(x_d = 1 \mid y = 1)$.

How would you estimate these parameters?

PARAMETER ESTIMATION

Reasonable way to set parameters:

- Set $p(y = 0)$ and $p(y = 1)$ to the empirical fraction of not spam/spam emails.
- For each word i, set $p(x_i = 1 \mid y = 0)$ to the empirical probability word i appears in a non-spam email.
- For each word i, set $\underline{p(x_i = 1 \mid y = 1)}$ to the empirical probability word i appears in a spam email.

Estimating these parameters is the only “training” we will do.

DONE WITH MODELING

ON TO PREDICTION

CLASSIFICATION RULE

Given unlabeled input $(\mathbf{x}, \_\_\_)$, choose the label $y \in \{0, 1\}$ which is most likely given the data. Recall $\mathbf{x} = [0, 0, 1, \ldots, 1, 0]$.

Classification rule: maximum a posterior prob. (MAP) estimate.

Step 1. Compute:

- $p(y = 0 \mid \mathbf{x})$: prob. $y = 0$ given observed data vector $\mathbf{x}$.
- $p(y = 1 \mid \mathbf{x})$: prob. $y = 1$ given observed data vector $\mathbf{x}$.

Step 2. Output: 0 or 1 depending on which probability is larger.

$p(y = 0 \mid \mathbf{x})$ and $p(y = 1 \mid \mathbf{x})$ are called **posterior** probabilities.

EVALUATING THE POSTERIOR

How to compute the posterior? **Bayes rule!**

$$\underline{p(y = 0 \mid \mathrm{x})} = \frac{p(\mathrm{x} \mid y = 0)p(y = 0)}{p(\mathrm{x})} \tag{1}$$

$$p(y=1 \mid x) = p(x \mid y=1)\, p(y=1) \,/\, p(x)$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \tag{2}$$

- **Prior:** Probability in class 0 prior to seeing any data.
- **Posterior:** Probability in class 0 after seeing the data.

EVALUATING THE POSTERIOR

Goal is to determine which is larger:

$$p(y=0 \mid \mathbf{x}) = \frac{\boxed{p(\mathbf{x} \mid y=0)}\boxed{p(y=0)}}{\cancel{p(\mathbf{x})}} \qquad \text{vs.}$$

$$p(y=1 \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y=1)\boxed{p(y=1)}}{\cancel{p(\mathbf{x})}}$$

How to compute posteriors:

- Ignore evidence $p(\mathbf{x})$ since it is the same for both sides.
- $p(y=0)$ and $p(y=1)$ already known (computed from training data).
- $p(\mathbf{x} \mid y=0) = ?$ $p(\mathbf{x} \mid y=1) = ?$

NAIVE BAYES

“Naive” Bayes Rule: Compute $p(\mathbf{x} \mid y = 0)$ by assuming independence:

model parameters chosen using data

$$p(\mathbf{x} \mid y = 0) = p(x_1 \mid y = 0) \cdot p(x_2 \mid y = 0) \cdot \ldots \cdot p(x_d \mid y = 0)$$

$\{0, 1, 1, 0, 0, \ldots\}$ → 0, 1

- $p(x_i \mid y = 0)$ is the probability you observe x_i given that an email is not spam.[1]

A more complicated method might take dependencies into account.

[1] Recall, x_i is either 0 when word i is not present, or 1 when word i is present.

Final Naive Bayes Classifier

Training/Modeling: Use existing data to compute:

- $p(y = 0), p(y = 1)$
- For all i:
    - Compute $p(0 \text{ at position } i \mid y = 0), p(1 \text{ at position } i \mid y_0)$
    - Compute $p(0 \text{ at position } i \mid y = 1), p(1 \text{ at position } i \mid y = 1)$

Prediction:

- For all i:
    - Compute $p(\mathbf{x} \mid y = 0) = \prod_i p(x_i \mid y = 0)$
    - Compute $p(\mathbf{x} \mid y = 1) = \prod_i p(x_i \mid y = 1)$
- Return

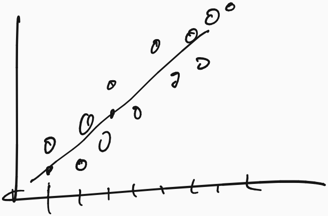
$$\arg\max \left[p\left(\mathbf{x} \mid y = 0\right) \cdot p\left(y = 0\right), p\left(\mathbf{x} \mid y = 1\right) \cdot p\left(y = 1\right)\right].$$

OTHER APPLICATIONS OF

THE BAYESIAN PERSPECTIVE

BAYESIAN REGRESSION

The Bayesian view offers an interesting alternative perspective on many machine learning techniques.

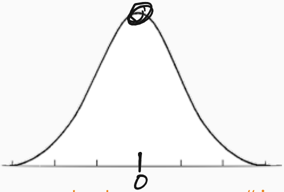


Example: Linear Regression.

Probabilistic model:

$$y_i = \underline{\langle \mathbf{x}_i, \beta \rangle} + \boxed{\eta}$$

where the η drawn from $N(0, \sigma^2)$ is **random Gaussian noise**.



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

$$Pr(\eta) = e^{-\eta^2/\sigma^2}$$

The symbol $\sim$ means “is proportional to”.

QUICK CHECK

$$y^* = \arg\max_{y} \; p(y \mid x_i)$$

Example: Linear Regression.

Probabilistic model:

$$p(y \mid x_i) = e^{-\frac{(\overbrace{y - \langle x_i, \beta \rangle}^{\eta})^2}{\sigma^2}}$$

$$y_i = \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle + \eta$$

where the η drawn from $N(0, \sigma^2)$ is **random Gaussian noise**.

If we knew $\boldsymbol{\beta}$ what is the maximum a posterior (MAP) estimate for y_i given observed data $\mathbf{x}_i$?

$$y^* = \langle x_i, \beta \rangle \qquad e^{-0^2/\sigma^2} \qquad 1$$

BAYESIAN REGRESSION

How should be select $\boldsymbol{\beta}$ for our model?

Bayesian approach: Use MAP estimate again! Now for parameter vector.

Choose $\boldsymbol{\beta}$ to maximize:

$$\Pr(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}) = \frac{\Pr(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\beta}) \Pr(\boldsymbol{\beta})}{\Pr(\mathbf{X}, \mathbf{y})}.$$

Assume all $\boldsymbol{\beta}$'s are equally likely, so we only care about $\Pr(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\beta})$ when maximizing.

Choose $\boldsymbol{\beta}$ to maximize:

$$\Pr(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\beta}) \sim$$

BAYESIAN REGRESSION

- $y_i = \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle + \eta$
- where $p(\eta = z) \sim e^{-z^2/\sigma^2}$

$$\Pr(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\beta}) \sim$$

LOG LIKELIHOOD

Easier to work with the log likelihood:

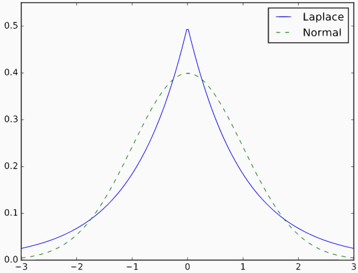
$$\begin{aligned}
\arg\max_{\boldsymbol{\beta}} \Pr(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\beta}) &= \arg\max_{\boldsymbol{\beta}} \prod_{i=1}^{n} e^{-(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2/\sigma^2} \\
&= \arg\max_{\boldsymbol{\beta}} \; \log\left(\prod_{i=1}^{n} e^{-(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2/\sigma^2} \right) \\
&= \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} -(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2/\sigma^2 \\
&= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2.
\end{aligned}$$

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

This is a completely different justification for squared loss.

BAYESIAN REGRESSION

If we had modeled our noise η as Laplace noise, we would have found that minimizing $\|\mathbf{y} - \mathbf{X}\beta\|_1$ was optimal.



$$Pr(\eta = z) \sim$$

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

This is a completely different justification for ℓ_1 loss.

BAYESIAN REGULARIZATION

Recall goal is to maximize over β:

$$\Pr(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}) = \frac{\Pr(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\beta}) \Pr(\boldsymbol{\beta})}{\Pr(\mathbf{X}, \mathbf{y})}.$$

~~assume all $\boldsymbol{\beta}$'s equally likely~~

Bayesian view: Assume values in $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_d]$ come from some distribution.

- **Common model:** Each β_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 $\boldsymbol{\beta}$ to maximize:

$$\Pr(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}) = \frac{\Pr(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\beta}) \Pr(\boldsymbol{\beta})}{\Pr(\mathbf{X}, \mathbf{y})}.$$

- We can still ignore the "evidence" term $\Pr(\mathbf{X}, \mathbf{y})$ since it is a constant that does not depend on $\boldsymbol{\beta}$.
- $\Pr(\boldsymbol{\beta}) = \Pr(\beta_1) \cdot \Pr(\beta_2) \cdot \ldots \cdot \Pr(\beta_d)$
- $\Pr(\boldsymbol{\beta}) \sim$

BAYESIAN REGULARIZATION

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

$$\begin{aligned}
&\arg\max_{\boldsymbol{\beta}} \Pr(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\beta}) \cdot \Pr(\boldsymbol{\beta}) \\
&= \arg\max_{\boldsymbol{\beta}} \prod_{i=1}^{n} e^{-(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2/\sigma^2} \cdot \prod_{i=1}^{n} e^{-(\beta_i)^2/\gamma^2} \\
&= \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} -(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2/\sigma^2 + \sum_{i=1}^{d} -(\beta_i)^2/\gamma^2 \\
&= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2 + \frac{\sigma^2}{\gamma^2} \sum_{i=1}^{d} (\beta_i)^2/\sigma^2.
\end{aligned}$$

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

Completely different justification for ridge regularization!

Test your intuition: What modeling assumption justifies LASSO regularization: $\min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$?