CSC 311: Introduction to Machine Learning

Lecture 7 - Probabilistic Models

Rahul G. Krishnan   Alice Gao

University of Toronto, Fall 2022

Outline

1. Probabilistic Modeling of Data
2. Discriminative and Generative Classifiers
3. Naïve Bayes Models
4. Bayesian Parameter Estimation
5. Multivariate Gaussian Distribution

Today

- So far in the course we have adopted a modular perspective, in which the model, loss function, optimizer, and regularizer are specified separately.
- Today we begin putting together a probabilistic interpretation of our model and loss, and introduce the concept of maximum likelihood estimation.

1 Probabilistic Modeling of Data

2 Discriminative and Generative Classifiers

3 Naïve Bayes Models

4 Bayesian Parameter Estimation

5 Multivariate Gaussian Distribution

Example: A Biased Coin

You flip a coin $N = 100$ times and get outcomes $\{x_1, \ldots, x_N\}$ where $x_i \in \{0, 1\}$ and $x_i = 1$ is interpreted as heads H.

Suppose you had $N_H = 55$ heads and $N_T = 45$ tails.

We want to create a model to predict the outcome of the next coin flip. That is, we want to answer this question:

What is the probability it will come up heads if we flip again?

Model

a discrete prob. dist. takes value 1 w/ prob θ
takes value 0 w/ prob $(1-\theta)$.

The coin is likely biased. Let's assume that one coin flip outcome x is a Bernoulli random variable for *a currently unknown parameter* $\theta \in [0, 1]$.

$$p(x = 1|\theta) = \theta \quad \text{and} \quad p(x = 0|\theta) = 1 - \theta$$

$$\text{or more succinctly} \quad p(x|\theta) = \theta^x (1 - \theta)^{1-x}$$

Assume that $\{x_1, \ldots, x_N\}$ are independent and identically distributed (i.i.d.). Thus, the joint probability of the outcome $\{x_1, \ldots, x_N\}$ is

$$p(x_1, ..., x_N|\theta) = \prod_{i=1}^{N} \theta^{x_i} (1 - \theta)^{1-x_i}$$

Loss Function

The likelihood function is the probability of observing the data as a function of the parameters θ: 55 heads, 45 tails

$$p(x_1, \ldots, x_N | \theta) = L(\theta) = \prod_{i=1}^{N} \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{55} (1-\theta)^{45}$$

We usually work with log-likelihoods:

$$\log p(x_1, \ldots, x_N | \theta) = \ell(\theta) = \sum_{i=1}^{N} x_i \log \theta + (1-x_i) \log(1-\theta)$$

$$= 55 \log \theta + 45 \log(1-\theta)$$

Maximum Likelihood Estimation

How can we choose θ? Good values of θ should assign high probability to the observed data.

The maximum likelihood criterion says that we should pick the parameters that maximize the likelihood. of data given parameters.

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta \in [0,1]} \ell(\theta)$$

We can find the optimal solution by setting derivatives to zero.

$$\frac{\mathrm{d}\ell}{\mathrm{d}\theta} = \frac{\mathrm{d}}{\mathrm{d}\theta}\left(\sum_{i=1}^{N} x_i \log\theta + (1 - x_i)\log(1-\theta)\right) = \frac{N_H}{\theta} - \frac{N_T}{1-\theta}$$

where $N_H = \sum_i x_i$ and $N_T = N - \sum_i x_i$.

Setting this to zero gives the maximum likelihood estimate:

$$\hat{\theta}_{\mathrm{ML}} = \frac{N_H}{N_H + N_T}. = \frac{55}{55+45} = 0.55$$

Maximum Likelihood Estimation

- define a model that assigns a probability (or has a probability density at) to a dataset
- maximize the likelihood (or minimize the neg. log-likelihood).

Observe N outcomes of the coin flip $\{x_1, \ldots, x_N\}$

$x_i \in \{0, 1\}$ $x_i = 1$ means heads. (H).

55 heads ($N_H = 55$), 45 tails ($N_T = 45$).

θ is the probability of the coin landing on heads.

$$\Pr(x_i \mid \theta) = \theta^{x_i} (1-\theta)^{1-x_i} = \begin{cases} \theta, & \text{if } x_i = 1. \\ 1-\theta, & \text{if } x_i = 0. \end{cases}$$

$$\Pr(x_1, \ldots, x_N \mid \theta)$$

$$\parallel$$

$$L(\theta) = \Pr(x_1, x_2, \ldots, x_N \mid \theta) = \prod_{i=1}^{N} \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{55} (1-\theta)^{45}$$

$$
\begin{aligned}
\ell(\theta) = \log \Pr(x_1, x_2, \ldots, x_N \mid \theta) &= \log \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i} \\
&= \sum_{i=1}^{N} \log\left(\theta^{x_i}(1-\theta)^{1-x_i}\right) \\
&= \sum_{i=1}^{N} \left(\log \theta^{x_i} + \log(1-\theta)^{1-x_i}\right) \\
&= \sum_{i=1}^{N} \left(x_i \log \theta + (1-x_i)\log(1-\theta)\right) \\
&= N_H \log \theta + N_T \log(1-\theta). \\
&= 55 \log \theta + 45 \log(1-\theta)
\end{aligned}
$$

$$\hat{\theta}_{\substack{\text{maximum}\\\text{likelihood}}} = \underset{\theta \in [0,1]}{\arg\max}\ \ell(\theta)$$

$$\frac{d\,\ell(\theta)}{d\theta} = \frac{d}{d\theta} \sum_{i=1}^{N} \left(x_i \log\theta + (1-x_i)\log(1-\theta) \right)$$

$$= \sum_{i=1}^{N} \left(\frac{x_i}{\theta} - \frac{1-x_i}{1-\theta} \right)$$

$$= \frac{N_H}{\theta} - \frac{N_T}{1-\theta} = \frac{55}{\theta} - \frac{45}{1-\theta}$$

$$\frac{d\,\ell(\theta)}{d\theta} = \frac{N_H}{\theta} - \frac{N_T}{1-\theta} = 0 \Longrightarrow \frac{N_H - N_H\theta - N_T\theta}{\theta(1-\theta)} = 0$$

$$N_H = \theta(N_H + N_T) \Longrightarrow \theta = \frac{N_H}{N_H+N_T} = \frac{55}{55+45} = 0.55$$

Summary of Maximum Likelihood:

~ model parameters θ. some data D.

~ calculate the log-likelihood of data given model parameters.

$$\log p(D|\theta)$$

~ choose model parameters that maximizes the log-likelihood.

$$\hat{\theta}_{ML} = \arg\max_{\theta} \log p(D|\theta)$$

For coin flip example.

$$\hat{\theta}_{ML} \text{ (prob of heads)} = \frac{\text{\# of heads}}{\text{\# of coin flips}}.$$

1 Probabilistic Modeling of Data

2 Discriminative and Generative Classifiers

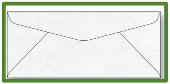
3 Naïve Bayes Models

4 Bayesian Parameter Estimation

5 Multivariate Gaussian Distribution

Spam Classification

For a large company that runs an email service, one of the important predictive problems is the automated detection of spam email.

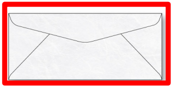


Dear Karim,

I think we should postpone the board meeting to be held after Thanksgiving.

Regards,
Anna

Not spam



Dear Toby,

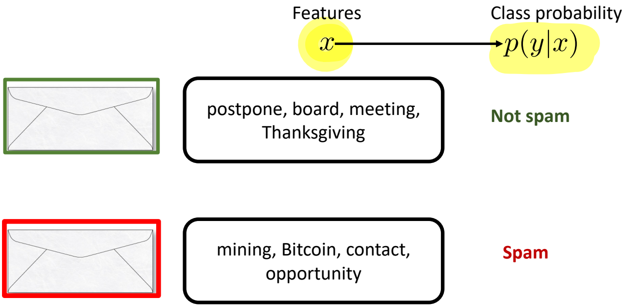
I have an incredible opportunity for mining 2 Bitcoin a day. Please Contact me at the earliest at +1 123 321 1555. You won’t want to miss out on this opportunity.

Regards,
Ark

Spam

Discriminative Classifiers

Discriminative classifiers try to learn mappings directly from the space of inputs $\mathcal{X}$ to class labels $\{0, 1, 2, \ldots, K\}$



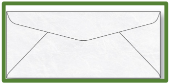
Generative Classifiers

Generative classifiers try to build a model of "what data for a class looks like", i.e. model $p(\mathbf{x}, y)$. If we know $p(y)$ we can easily compute $p(\mathbf{x}|y)$.

Classification via Bayes rule (thus also called Bayes classifiers)

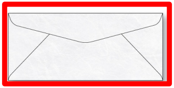
$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Probability of feature given label $p(x|y)$ ← Class label y



postpone, board, meeting, Thanksgiving

Not spam



mining, Bitcoin, contact, opportunity

Spam

Generative vs Discriminative

- Discriminative approach: estimate parameters of decision boundary/class separator directly from labeled examples.
  - Model $p(t|\mathbf{x})$ directly (logistic regression models)
  - Learn mappings from inputs to classes (linear/logistic regression, decision trees etc)
  - Tries to solve: How do I separate the classes?
- Generative approach: model the distribution of inputs characteristic of the class (Bayes classifier).
  - Model $p(\mathbf{x}|t)$
  - Apply Bayes Rule to derive $p(t|\mathbf{x})$.
  - Tries to solve: What does each class "look" like?
- Key difference: is there a distributional assumption over inputs?

1. Probabilistic Modeling of Data
2. Discriminative and Generative Classifiers
3. Naïve Bayes Models
4. Bayesian Parameter Estimation
5. Multivariate Gaussian Distribution

Example: Spam Detection

- Classify email into spam ($c = 1$) or non-spam ($c = 0$).
- Binary features $\mathbf{x} = [x_1, \ldots, x_D]$, $x_i \in \{0, 1\}$ saying whether each of D words appears in the e-mail.

Example email: "You are one of the very few who have been selected as a winner for the free $1000 Gift Card."

Feature vector for this email:

- ...
- "card": 1
- ...
- "winners": 1
- "winter": 0
- ...
- "you": 1

Bayesian Classifier

Given features $\mathbf{x} = [x_1, x_2, \cdots, x_D]^T$
want to compute class probabilities using Bayes Rule:

$$\underbrace{p(c|\mathbf{x})}_{\text{Pr. class given feature}} = \frac{\overbrace{p(\mathbf{x}|c)}^{\text{Pr. feature given class}} \quad p(c)}{p(\mathbf{x})}$$

In words,

$$\text{Posterior for class} = \frac{\text{Pr. of feature given class} \times \text{Prior for class}}{\text{Pr. of feature}}$$

To compute $p(c|\mathbf{x})$ we need: $p(\mathbf{x}|c)$ and $p(c)$.

① explain each term

② $p(x|c) \rightarrow p(c|x)$

③ prior → posterior

$$\Pr\left(\begin{matrix}\text{word in}\\ \text{an email}\end{matrix} \middle| \begin{matrix}\text{spam}\\ \text{or not}\end{matrix}\right)$$

$\Pr(\text{spam})$

$\Pr(\text{non-spam})$

Prior

$$\Pr(c|x) = \frac{\Pr(x|c) \quad \Pr(c)}{\Pr(x)}$$

$$\Pr\left(\begin{matrix}\text{spam}\\ \text{or not}\end{matrix} \middle| \begin{matrix}\text{word in}\\ \text{an email}\end{matrix}\right)$$

Posterior

$\Pr(\text{word in an email})$

e.g. $\Pr(\text{"winner"})$

$\Pr(\text{"you"})$

Do not need $p(x)$ explicitly. It's a normalization constant.

$$p(C=1|x) = \frac{p(x|c=1)\,p(C=1)}{p(x)}$$

$$= \frac{p(x|c=1)\,p(C=1)}{p(x|c=1)\,p(C=1) + p(x|c=0)\,p(C=0)}$$

① Calculate $p(x|c=1)\,p(c=1)$ and $p(x|c=0)\,p(c=0)$.

② then normalize. (divide each by the sum of the two.)

Motivation for Compact Representation

- Two classes: $c \in \{0, 1\}$.
- Binary features $\mathbf{x} = [x_1, \dots, x_D], x_i \in \{0, 1\}$
- Define a joint distribution $p(c, x_1, \dots, x_D)$.
  How many probabilities do we need to specify this joint dist.?
  $2^{D+1} - 1$
- Let's impose structure on the distribution so that
  the representation is compact and
  allows for efficient learning and inference

Naïve Bayes Independence Assumption

Naïve assumption:
the features x_i are conditionally independent given the class c.

- Allows us to decompose the joint distribution:

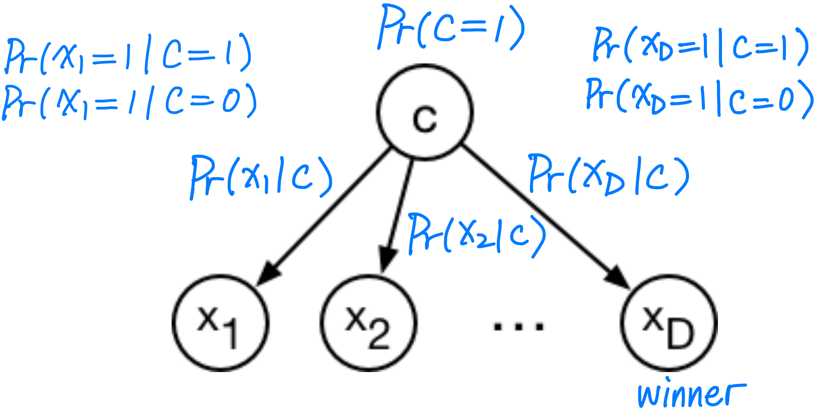
$$p(c, x_1, \ldots, x_D) = \underbrace{p(c)}_{\pi}\, \underbrace{p(x_1|c)}_{\theta_{1c}} \cdots \underbrace{p(x_D|c)}_{\theta_{Dc}}.$$

Compact representation of the joint distribution

- Prior probability of class:
  $p(c = 1) = \pi$ (e.g. prob of spam)

- Conditional probability of feature given class:
  $p(x_j = 1|c) = \theta_{jc}$ (e.g. prob of word appearing in spam)

Bayesian Network for a Naive Bayes Model



- Which probabilities do we need to specify this dist.?
- How many probabilities do we need to specify this dist.?

$1 + 2D$

Decomposing the Log-Likelihood

Decompose the log-likelihood into independent terms.
Optimize each term independently.

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log p(c^{(i)}, \mathbf{x}^{(i)}) = \sum_{i=1}^{N} \log \left\{ p(\mathbf{x}^{(i)} | c^{(i)}) p(c^{(i)}) \right\} \\
&= \sum_{i=1}^{N} \log \left\{ p(c^{(i)}) \prod_{j=1}^{D} p(x_j^{(i)} \mid c^{(i)}) \right\} \\
&= \sum_{i=1}^{N} \left[\log p(c^{(i)}) + \sum_{j=1}^{D} \log p(x_j^{(i)} \mid c^{(i)}) \right] \\
&= \underbrace{\sum_{i=1}^{N} \log p(c^{(i)})}_{\substack{\text{Log-likelihood} \\ \text{of labels}}} + \sum_{j=1}^{D} \underbrace{\sum_{i=1}^{N} \log p(x_j^{(i)} \mid c^{(i)})}_{\substack{\text{Log-likelihood} \\ \text{for feature } x_j}}
\end{aligned}$$

Learning the Prior over Class

- To learn the prior, we maximize $\sum_{i=1}^{N} \log p(c^{(i)})$
- Define $\pi = p(c^{(i)} = 1)$
- Pr. i-th email: $p(c^{(i)}) = \pi^{c^{(i)}}(1-\pi)^{1-c^{(i)}}$.
- Log-likelihood of the dataset:

$$\sum_{i=1}^{N} \log p(c^{(i)}) = \sum_{i=1}^{N} c^{(i)} \log \pi + \sum_{i=1}^{N} (1 - c^{(i)}) \log(1 - \pi)$$

- Maximum likelihood estimate of the prior π is the fraction of spams in dataset.

$$\hat{\pi} = \frac{\sum_i \mathbb{1}[c^{(i)} = 1]}{N} = \frac{\#\text{ spams in dataset}}{\text{total }\#\text{ samples}}$$

$c^{(i)} \in \{0, 1\}$ is the class label for i^{th} example.

$$p(c^{(i)} | \pi) = \pi^{c^{(i)}} (1-\pi)^{1-c^{(i)}}$$

$$p(c^{(1)}, c^{(2)}, \dots, c^{(N)} | \pi) = \prod_{i=1}^{N} \pi^{c^{(i)}} (1-\pi)^{1-c^{(i)}}$$

$$\log p(c^{(1)}, \dots, c^{(N)} | \pi) = \log \prod_{i=1}^{N} \pi^{c^{(i)}} (1-\pi)^{1-c^{(i)}}$$

$$= \sum_{i=1}^{N} \log\left(\pi^{c^{(i)}} (1-\pi)^{1-c^{(i)}}\right)$$

$$= \sum_{i=1}^{N} \left(c^{(i)} \log \pi + (1-c^{(i)}) \log (1-\pi)\right)$$

$$\frac{\partial}{\partial \pi} \log p(c^{(1)}, \dots, c^{(N)} | \pi) = \sum_{i=1}^{N} \left(c^{(i)} \frac{1}{\pi} - (1-c^{(i)}) \frac{1}{1-\pi}\right)$$

Let $\sum_{i=1}^{N} \mathbb{I}[c^{(i)}=1] = S$

$$\frac{\partial}{\partial \pi} \log p(c^{(1)}, \dots, c^{(N)} \mid \pi) = \sum_{i=1}^{N} \left(c^{(i)} \frac{1}{\pi} - (1-c^{(i)}) \frac{1}{1-\pi} \right)$$

$$= \frac{S}{\pi} - \frac{N-S}{1-\pi} = 0$$

$$\frac{S}{\pi} = \frac{N-S}{1-\pi} \implies S(1-\pi) = (N-S)\pi$$

$$\implies S = S\pi + (N-S)\pi$$

$$\implies S = N\pi$$

$$\implies \pi = \frac{S}{N} = \frac{\sum_{i=1}^{N} \mathbb{I}[c^{(i)}=1]}{N}$$

Learning Pr. Feature Given Class

- To learn $p(x_j^{(i)} = 1 \,|\, c)$, we maximize $\sum_{i=1}^{N} \log p(x_j^{(i)} | \, c^{(i)})$
- Define $\theta_{jc} = p(x_j^{(i)} = 1 \,|\, c)$.
- Pr. of i-th email: $p(x_j^{(i)} \,|\, c) = \theta_{jc}^{x_j^{(i)}} (1 - \theta_{jc})^{1 - x_j^{(i)}}$.
- Log-likelihood of the dataset:

$$\sum_{i=1}^{N} \log p(x_j^{(i)} \,|\, c^{(i)}) = \sum_{i=1}^{N} c^{(i)} \left\{ x_j^{(i)} \log \theta_{j1} + (1 - x_j^{(i)}) \log(1 - \theta_{j1}) \right\}$$

$$+ \sum_{i=1}^{N} (1 - c^{(i)}) \left\{ x_j^{(i)} \log \theta_{j0} + (1 - x_j^{(i)}) \log(1 - \theta_{j0}) \right\}$$

- Maximum likelihood estimate of θ_{jc} is the fraction of word j occurrances in each class in the dataset.

$$\hat{\theta}_{jc} = \frac{\sum_i \mathbb{1}[x_j^{(i)} = 1 \;\&\; c^{(i)} = c]}{\sum_i \mathbb{1}[c^{(i)} = c]} \overset{\text{for } c = 1}{=} \frac{\#\text{word } j \text{ appears in class } c}{\#\text{ class } c \text{ in dataset}}$$

$x_j^{(i)} \in \{0,1\}$ denotes whether jth word in dictionary occurs in ith email.

$$p(x_j^{(i)} | c^{(i)}) = p(x_j^{(i)} | c^{(i)})^{x_j^{(i)}} (1 - p(x_j^{(i)} | c^{(i)}))^{1 - x_j^{(i)}}$$

$$\log p(x_j^{(i)} | c^{(i)}) = x_j^{(i)} \log P(x_j^{(i)} = 1 | c^{(i)}) + (1 - x_j^{(i)}) \log (1 - P(x_j^{(i)} = 1 | c^{(i)}))$$

$$\sum_{i=1}^{N} \log p(x_j^{(i)} | c^{(i)}) = \sum_{i=1}^{N} \left[x_j^{(i)} \log P(x_j^{(i)} = 1 | c^{(i)}) + (1 - x_j^{(i)}) \log (1 - P(x_j^{(i)} = 1 | c^{(i)})) \right]$$

$$= \sum_{i=1}^{N} c^{(i)} \left[x_j^{(i)} \log P(x_j^{(i)} = 1 | c^{(i)} = 1) + (1 - x_j^{(i)}) \log (1 - P(x_j^{(i)} = 1 | c^{(i)} = 1)) \right]$$

$$+ \sum_{i=1}^{N} (1 - c^{(i)}) \left[x_j^{(i)} \log P(x_j^{(i)} = 1 | c^{(i)} = 0) + (1 - x_j^{(i)}) \log (1 - P(x_j^{(i)} = 1 | c^{(i)} = 0)) \right]$$

$$\frac{\partial \sum_{i=1}^{N} \log P(x_j^{(i)} \mid c^{(i)})}{\partial P(x_j^{(i)}=1 \mid c^{(i)}=1)} = \sum_{i=1}^{N}\left[c^{(i)}\left(\frac{x_j^{(i)}}{P(x_j^{(i)}=1 \mid c^{(i)}=1)} - \frac{1-x_j^{(i)}}{P(x_j^{(i)}=1 \mid c^{(i)}=1)}\right)\right] = 0$$

$$\boxed{\text{Let } \theta_{j1} = P(x_j^{(i)}=1 \mid c^{(i)}=1)}$$

$$\Rightarrow \sum_{i=1}^{N} c^{(i)}\left(x_j^{(i)}(1-\theta_{j1}) - (1-x_j^{(i)})\,\theta_{j1}\right) = 0$$

$$\Rightarrow \sum_{i=1}^{N} c^{(i)}\left(x_j^{(i)} - \theta_{j1}\right) = 0$$

$$\Rightarrow \sum_{i=1}^{N} c^{(i)} x_j^{(i)} = \theta_{j1} \sum_{i=1}^{N} c^{(i)} \Rightarrow \theta_{j1} = \frac{\sum_{i=1}^{N} c^{(i)} x_j^{(i)}}{\sum_{i=1}^{N} c^{(i)}}$$

Predicting the Most Likely Class

- We predict the class by performing inference in the model.
- Apply Bayes' Rule:

$$p(c \mid \mathbf{x}) = \frac{p(c)p(\mathbf{x} \mid c)}{\sum_{c'} p(c')p(\mathbf{x} \mid c')} = \frac{p(c) \prod_{j=1}^{D} p(x_j \mid c)}{\cancel{\sum_{c'} p(c') \prod_{j=1}^{D} p(x_j \mid c')}}$$

- For input $\mathbf{x}$, predict c with the largest $p(c) \prod_{j=1}^{D} p(x_j \mid c)$ (the most likely class).

proportional to

$$p(c \mid \mathbf{x}) \propto p(c) \prod_{j=1}^{D} p(x_j \mid c)$$

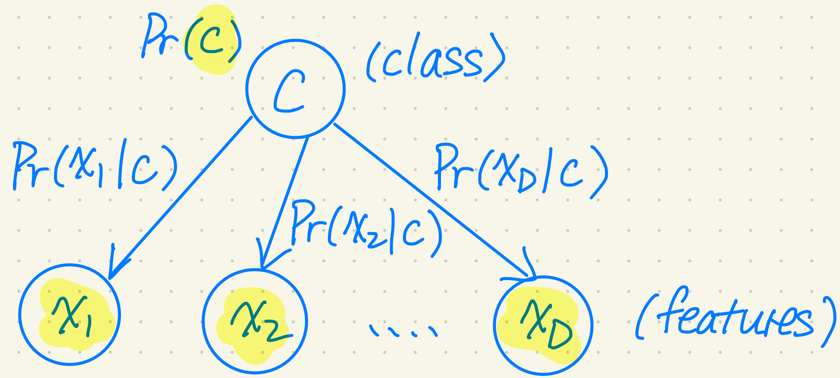
Naïve Bayes Properties

- An amazingly cheap learning algorithm!
- Training time: estimate parameters using maximum likelihood
  - Compute co-occurrence counts of each feature with the labels.
  - Requires only one pass through the data!
- Test time: apply Bayes' Rule
  - Cheap because of the model structure. (For more general models, Bayesian inference can be very expensive and/or complicated.)
- Analysis easily extends to prob. distributions other than Bernoulli.
- Less accurate in practice compared to discriminative models due to its "naïve" independence assumption.

Naive Bayes Summary.

Model Parameters:

$$\begin{cases} Pr(c) = \pi \\ Pr(x_j | c) = \theta_{jc} \end{cases}$$



① Learning the model parameters.

~ Learn π by maximum likelihood.

~ Learn θ_{jc} by maximum likelihood.

② Making a prediction.

for input x, predict class c w/ largest $p(c|x) \propto p(c) \prod_{j=1}^{D} p(x_j|c)$

1 Probabilistic Modeling of Data

2 Discriminative and Generative Classifiers

3 Naïve Bayes Models

4 Bayesian Parameter Estimation

5 Multivariate Gaussian Distribution

Data Sparsity

Maximum likelihood can overfit if there is too little data.

Example: what if you flip the coin twice and get H both times?

$$\theta_{\mathrm{ML}} = \frac{N_H}{N_H + N_T} = \frac{2}{2+0} = 1$$

The model assigned probability 0 to T.
This problem is known as data sparsity.

Defining a Bayesian Model

We need to specify two distributions:

- The prior distribution $p(\boldsymbol{\theta})$
  encodes our beliefs about the parameters *before* we observe the data.

- The likelihood $p(\mathcal{D} \mid \boldsymbol{\theta})$
  encodes the likelihood of observing the data given the parameters.

The Posterior Distribution

- When we update our beliefs based on the observations, we compute the posterior distribution using Bayes' Rule:

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\boldsymbol{\theta}) p(\mathcal{D} \mid \boldsymbol{\theta})}{\int p(\boldsymbol{\theta}') p(\mathcal{D} \mid \boldsymbol{\theta}') \, \mathrm{d}\boldsymbol{\theta}'}.$$

- Rarely ever compute the denominator explicitly.
- In general, computing the denominator is intractable.

Revisiting Coin Flip Example

We already know the likelihood:

$$L(\theta) = p(\mathcal{D}|\theta) = \theta^{N_H}(1-\theta)^{N_T}$$

It remains to specify the prior $p(\theta)$.

- An uninformative prior, which assumes as little as possible. A reasonable choice is the uniform prior.
- But, experience tells us 0.5 is more likely than 0.99. One particularly useful prior is the beta distribution:

$$p(\theta; a, b) = \boxed{\cancel{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}}\,\theta^{a-1}(1-\theta)^{b-1}.$$

- We can ignore $\boxed{\text{the normalization constant.}}$

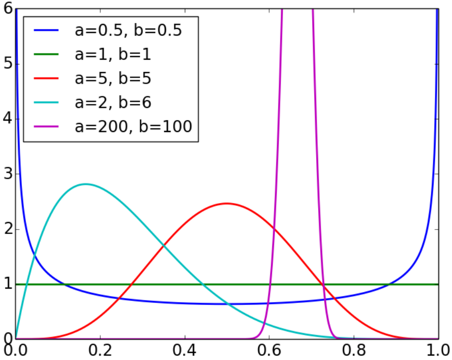
$$p(\theta; a, b) \propto \theta^{a-1}(1-\theta)^{b-1}.$$

proportional to

Beta Distribution Properties

- The expectation is $\mathbb{E}[\theta] = a/(a+b)$. a=b symmetric about 0.5.
- The distribution gets more peaked when a and b are large.
- When $a = b = 1$, it becomes the uniform distribution.

defined on [0,1].



Posterior for the Coin Flip Example

- Computing the posterior distribution: $P(\theta|D) = \dfrac{P(\theta)\,P(D|\theta)}{P(D)}$

prior

$P(\theta)$

$= \theta^{a-1}(1-\theta)^{b-1}$

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \mathcal{D}) &\propto p(\boldsymbol{\theta})p(\mathcal{D} \mid \boldsymbol{\theta}) \\
&\propto \left[\theta^{a-1}(1-\theta)^{b-1}\right]\left[\theta^{N_H}(1-\theta)^{N_T}\right] \\
&= \theta^{a-1+N_H}(1-\theta)^{b-1+N_T}.
\end{aligned}
$$

A beta distribution with parameters $N_H + a$ and $N_T + b$.

- The posterior expectation of θ is:

$$
\mathbb{E}[\theta \mid \mathcal{D}] = \frac{N_H + a}{N_H + N_T + a + b}
$$

For prior $E[\theta] = \dfrac{a}{a+b}$

- Think of a and b as pseudo-counts.
  beta(a, b) = $\boxed{\text{beta}(1, 1)}$ + $a - 1$ heads + $b - 1$ tails.

uniform prior

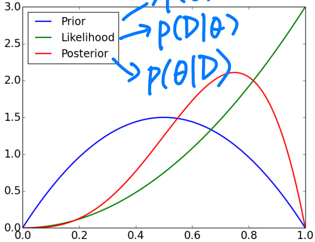
- The prior and likelihood have the same functional form (conjugate priors). prior & posterior in the same dist. family.

Bayesian Inference for the Coin Flip Example

When you have enough observations, the data overwhelm the prior.

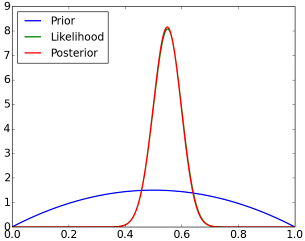
Small data setting
$N_H = 2,\ N_T = 0$

$\to p(\theta)$
$\to p(D|\theta)$
$\to p(\theta|D)$



$a-1+2$ heads
$b-1$ tails

Large data setting
$N_H = 55,\ N_T = 45$

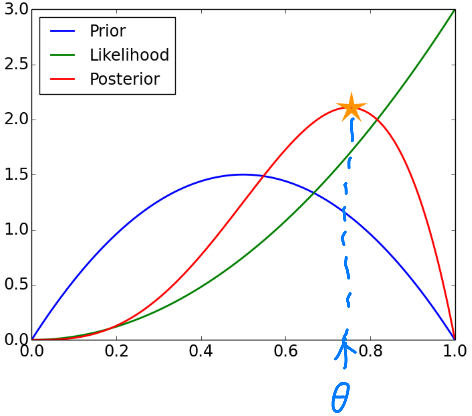


$a-1+55$ heads
$b-1+45$ tails.

Maximum A-Posteriori (MAP) Estimation

$P(\theta|D)$

Finds the most likely parameters under the posterior (i.e. the mode).



Maximum A-Posteriori Estimation

Converts the Bayesian parameter estimation problem into a maximization problem

if uniform prior, MAP = ML. since $p(\theta)$ is a constant.

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} &= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \mathcal{D}) \\
&= \arg\max_{\boldsymbol{\theta}} \boxed{p(\boldsymbol{\theta})}\, p(\mathcal{D} \mid \boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \underbrace{\log p(\boldsymbol{\theta})}_{\text{prior (like a regularizer)}} + \underbrace{\log p(\mathcal{D} \mid \boldsymbol{\theta})}_{\text{maximum likelihood.}}
\end{aligned}
$$

Maximum Likelihood.

$$\hat{\theta}_{ML} = \arg\max_{\theta} p(D \mid \theta)$$

Maximum A-Posteriori Estimation

$p(\theta)\, p(D|\theta)$

$$p(\theta|D) = \frac{p(\theta, D)}{p(D)}$$

Joint probability of parameters and data:

$= \log(p(\theta) * p(D|\theta))$

$$\begin{aligned}
\log p(\theta, \mathcal{D}) &= \log p(\theta) + \log p(\mathcal{D} \,|\, \theta) \\
&= \text{Const} + (N_H + a - 1) \log \theta + (N_T + b - 1) \log(1 - \theta)
\end{aligned}$$

Maximize by finding a critical point

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(\theta, \mathcal{D}) = \frac{N_H + a - 1}{\theta} - \frac{N_T + b - 1}{1 - \theta} = 0$$

Solving for θ,

$$\hat{\theta}_{\text{MAP}} = \frac{N_H + a - 1}{N_H + N_T + a + b - 2}$$

$a - 1 + N_H$ heads
$b - 1 + N_T$ tails.

Estimate Comparison for Coin Flip Example

overfitting.

	Formula	$N_H = 2, N_T = 0$	$N_H = 55, N_T = 45$
$\hat{\theta}_{\text{ML}}$	$\frac{N_H}{N_H+N_T}$	1	$\frac{55}{100} = 0.55$
$\mathbb{E}[\theta \vert \mathcal{D}]$	$\frac{N_H+a}{N_H+N_T+a+b}$	$\frac{4}{6} \approx 0.67$	$\frac{57}{104} \approx 0.548$
$\hat{\theta}_{\text{MAP}}$	$\frac{N_H+a-1}{N_H+N_T+a+b-2}$	$\frac{3}{4} = 0.75$	$\frac{56}{102} \approx 0.549$

$\hat{\theta}_{\text{MAP}}$ assigns nonzero probabilities as long as $a, b > 1$.

avoids overfitting.

Bayesian Parameter Estimation

- Maximum Likelihood overfits when there is little data.
- Add a prior (our belief before observing data).

$$\underbrace{P(\theta|D)}_{\text{posterior}} \propto \underbrace{P(\theta)}_{\text{prior}} \underbrace{P(D|\theta)}_{\text{likelihood}}.$$

- Maximum A Posteriori Estimation:

  choose model parameters that have the largest posterior probability.

$$\hat{\theta}_{MAP} = \arg\max_{\theta}\Big(\log P(\theta|D)\Big) = \arg\max_{\theta}\Big(\underbrace{\log P(\theta)}_{\substack{\text{prior}\\ \text{(regularizer)}}} + \underbrace{\log P(D|\theta)}_{\substack{\text{maximum}\\ \text{likelihood.}}}\Big)$$

Maximum A Posteriori Estimation for Coin Flip.

prior is the beta distribution.

$$p(\theta) = \theta^{a-1}(1-\theta)^{b-1} \qquad \log p(\theta) = (a-1)\log\theta + (b-1)\log(1-\theta).$$

Likelihood:

$$\log p(D|\theta) = N_H \log\theta + N_T \log(1-\theta)$$

posterior:

$$\log p(\theta|D) \propto \log p(\theta) + \log p(D|\theta)$$

$$= (N_H + a - 1)\log\theta + (N_T + b - 1)\log(1-\theta).$$

$$\hat{\theta}_{MAP} = \frac{N_H + a - 1}{(N_H + a - 1) + (N_T + b - 1)}$$

1 Probabilistic Modeling of Data

2 Discriminative and Generative Classifiers

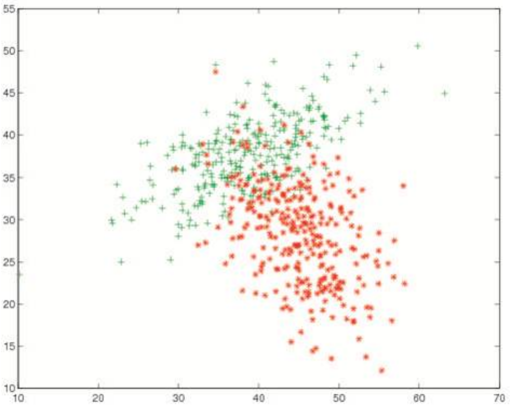
3 Naïve Bayes Models

4 Bayesian Parameter Estimation

5 Multivariate Gaussian Distribution

Classification: Diabetes Example

- Observation per patient: White blood cell count & glucose value.



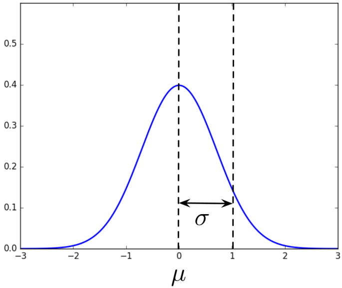
- $p(\mathbf{x} \mid t = k)$ for each class is shaped like an ellipse
  $\Longrightarrow$ we model each class as a multivariate Gaussian

Univariate Gaussian distribution

- Recall the Gaussian, or normal, distribution:

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Parameterized by mean μ and variance σ^2.
- The Central Limit Theorem says that sums of lots of independent random variables are approximately Gaussian.
- In machine learning, we use Gaussians a lot because they make the calculations easy.



Multivariate Mean and Covariance

- Mean

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}] = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix}$$

- Covariance

$$\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{x}) = \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1D} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{D1} & \sigma_{D2} & \cdots & \sigma_D^2 \end{pmatrix}$$

- The statistics ($\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) uniquely define a multivariate Gaussian (or multivariate Normal) distribution, denoted $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - This is not true for distributions in general!