Today’s topics

Bayes rule review

A probabilistic view of machine learning
  Joint Distributions
  Bayes optimal classifier

Statistical Estimation
  Maximum likelihood estimates
  Derive relative frequency as the solution to a constrained optimization problem
The Bayes Optimal Classifier

Assume we know the data generating distribution $\mathcal{D}$

We define the **Bayes Optimal classifier** as

$$f^{(BO)}(\hat{x}) = \arg \max_{\hat{y} \in \mathcal{Y}} \mathcal{D}(\hat{x}, \hat{y})$$

**Theorem**: Of all possible classifiers, the Bayes Optimal classifier achieves the smallest zero/one loss

**Bayes error rate**

Defined as the error rate of the Bayes optimal classifier

Best error rate we can ever hope to achieve under zero/one loss
The Bayes Optimal Classifier

Assume we know the data generating distribution $D$.

We define the **Bayes Optimal classifier** as

$$f^{(BO)}(\hat{x}) = \arg\max_{\hat{y}} D(\hat{x}, \hat{y})$$

**Theorem:** Of all possible classifiers, the Bayes Optimal classifier achieves the smallest zero/one loss, also known as the **Bayes error rate**.

If we had access to $D$, finding an optimal classifier would be trivial!

So let’s estimate it instead!
What does “training” mean in probabilistic settings?

- Training = estimating $\mathcal{D}$ from a finite training set
  - We typically assume that $\mathcal{D}$ comes from a specific family of probability distributions
  - e.g., Bernouilli, Gaussian, etc
  - Learning means inferring parameters of that distributions
  - e.g., mean and covariance of the Gaussian
Training assumption: training examples are iid

- Independently and Identically distributed
  - i.e. as we draw a sequence of examples from $\mathcal{D}$, the n-th draw is independent from the previous n-1 sample

- This assumption is usually false!
  - But sufficiently close to true to be useful
How can we estimate the joint probability distribution from data?
What are the challenges?
What we know so far...

- Bayes rule

- A probabilistic view of machine learning
  - If we know the data generating distribution, we can define the Bayes optimal classifier
  - Under iid assumption

- How to estimate a probability distribution from data?
  - Maximum likelihood estimation
Maximum Likelihood Estimation

- Find the parameters that maximize the probability of the data

- Example: how to model a biased coin? (on board)
Maximum Likelihood Estimates

Each coin flip yields a Boolean value for $X$

$$X \sim \text{Bernoulli} : P(X) = \theta^X (1 - \theta)^{1-X}$$

Given a data set $D$ of iid flips, which contains $\alpha_1$ ones and $\alpha_0$ zeros

$$P_\theta(D) = \theta^{\alpha_1} (1 - \theta)^{\alpha_0}$$

$$\hat{\theta}_{MLE} = \arg\max_\theta P_\theta(D) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$
Maximum Likelihood Estimation

- Exercise: how to model a k-sided die? (on board)
Let's learn a classifier by learning $P(Y|X)$

Goal: learn a classifier $P(Y|X)$

Prediction:
Given an example $x$
Predict $\hat{y} = \arg\max_y P(Y = y | X = x)$
Parameters for $P(X,Y)$ vs. $P(Y|X)$

$Y = \text{Wealth}$  
$X = <\text{Gender, Hours\_worked}>$

### Joint probability distribution $P(X,Y)$

<table>
<thead>
<tr>
<th>gender</th>
<th>hours_worked</th>
<th>wealth</th>
<th>$P(Y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>v0:40.5-</td>
<td>poor</td>
<td>0.253122</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0245895</td>
</tr>
<tr>
<td></td>
<td>v1:40.5+</td>
<td>poor</td>
<td>0.0421768</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0116293</td>
</tr>
<tr>
<td>Male</td>
<td>v0:40.5-</td>
<td>poor</td>
<td>0.331313</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0971295</td>
</tr>
<tr>
<td></td>
<td>v1:40.5+</td>
<td>poor</td>
<td>0.134106</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.105933</td>
</tr>
</tbody>
</table>

### Conditional probability distribution $P(Y|X)$

| Gender | HrsWorked | $P(\text{rich} | G,HW)$ | $P(\text{poor} | G,HW)$ |
|--------|-----------|----------|----------|
| F      | <40.5     | .09      | .91      |
| F      | >40.5     | .21      | .79      |
| M      | <40.5     | .23      | .77      |
| M      | >40.5     | .38      | .62      |
How many parameters do we need to learn?

Suppose $X = < X_1, X_2, ... X_d >$
where $X_i$ and $Y$ are Boolean random variables

Q: How many parameters do we need to estimate $P(Y|X_1, X_2, ... X_d)$?

A: Too many to estimate $P(Y|X)$ directly from data!
Naïve Bayes Assumption

Naïve Bayes assumes

\[ P(X_1, X_2, \ldots, X_d | Y) = \prod_{i=1}^{d} P(X_i | Y) \]

i.e., that \( X_i \) and \( X_j \) are conditionally independent given \( Y \), for all \( i \neq j \)
Conditional Independence

Definition:

X is conditionally independent of Y given Z if

\[ P(X|Y, Z) = P(X|Z) \]

Recall that X is independent of Y if

\[ P(X|Y) = P(X) \]
Naïve Bayes classifier

\[ \hat{y} = \arg\max_{y} P(Y = y | X = x) = \arg\max_{y} P(Y = y)P(X = x | Y = y) = \arg\max_{y} P(Y = y) \prod_{i=1}^{d} P(X_i = x_i | Y = y) \]

Bayes rule

+ Conditional independence assumption
How many parameters do we need to learn?

To describe $P(Y)$?

To describe $P(X = <X_1, X_2, \ldots, X_d> | Y)$

Without conditional independence assumption?

With conditional independence assumption?

(Suppose all random variables are Boolean)
Training a Naïve Bayes classifier

Let's assume discrete $X_i$ and $Y$

**TrainNaïveBayes** (Data)

for each value $y_k$ of $Y$

estimate $\pi_k = P(Y = y_k)$

for each value $x_{ij}$ of $X_i$

estimate $\theta_{ijk} = P(X_i = x_{ij} | Y = y_k)$

# examples for which $X_i = x_{ij}$ and $Y = y_k$

# examples for which $Y = y_k$

# examples for which $Y = y_k$

# examples
Naïve Bayes Wrap-up

An easy to implement classifier, that performs well in practice

Subtleties

Often the $X_i$ are not really conditionally independent
What if the Maximum Likelihood estimate for $P(X_i|Y)$ is zero?
What is the decision boundary of a Naïve Bayes classifier?
Naïve Bayes Properties

Naïve Bayes is a linear classifier

See CIML for example of computation of Log Likelihood Ratio

Choice of probability distribution is a form of inductive bias

99.7% of the data are within 3 standard deviations of the mean
95% within 2 standard deviations
68% within 1 standard deviation

FEARLESS IDEAS
Generative Stories

Probabilistic models tell a fictional story explaining how our training data was created

Example of a generative story for a multiclass classification task with continuous features

For each example $n = 1 \ldots N$:

(a) Choose a label $y_n \sim \text{Disc}(\theta)$

(b) For each feature $d = 1 \ldots D$:

i. Choose feature value $x_{n,d} \sim \text{Nor}(\mu_{y_n,d}, \sigma^2_{y_n,d})$
From the Generative Story to the Likelihood Function

For each example $n = 1 \ldots N$:

(a) Choose a label $y_n \sim \text{Disc}(\theta)$

(b) For each feature $d = 1 \ldots D$:
   
   i. Choose feature value $x_{n,d} \sim \mathcal{N}(\mu_{y_n,d}, \sigma^2_{y_n,d})$

\[
p(D) = \prod_n \theta_{y_n} \prod_d \frac{1}{\sqrt{2\pi\sigma^2_{y_n,d}}} \exp \left[ -\frac{1}{2\sigma^2_{y_n,d}} (x_{n,d} - \mu_{y_n,d})^2 \right]
\]
What you should know

The Naïve Bayes classifier
  Conditional independence assumption
  How to train it?
  How to make predictions?
  How does it relate to other classifiers we know?

Fundamental Machine Learning concepts
  iid assumption
  Bayes optimal classifier
  Maximum Likelihood estimation
  Generative story