Lecture 20: Naïve Bayes

Lisa Yan
August 10, 2018
Announcements

Final Exam Review:
◦ Survey for concepts/questions (please fill out by Monday 8/13): https://goo.gl/forms/UPsGbff3FY5MU5pf2

Note: Last OH on Wednesday 8/15 (None on Thursday 8/16)
Problem Set #6 coding question: covered today
Summary from last time

Parameter estimation:

Given a sample of size $n$ of IID RVs drawn from the same distribution:

- Find a “point estimate” of model parameters $\theta$
- (useful for simulating new samples later on)
- Intuitively, this should be the parameters that make the observed sample the most likely.
- Joint likelihood of seeing the particular sample $X_1, X_2, \ldots, X_n$:

$$L(\theta) = f(X_1, X_2, \ldots, X_n | \theta) = \prod_{i=1}^{n} f(X_i | \theta)$$

Likelihood function

Conditional densities are a function of the chosen parameter $\theta$
Maximum Likelihood Estimator

Consider the sample of $n$ IID RVs, $X_1, \ldots, X_n$, where $X_i$ has density function $f(X_i|\theta)$

$$L(\theta) = f(X_1, X_2, \ldots, X_n|\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

Likelihood function

Maximum Likelihood Estimator (MLE):

$$\theta_{MLE} = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} LL(\theta) = \arg \max_{\theta} \sum_{i=1}^{n} \log f(X_i|\theta)$$

1. Determine formula for log-likelihood $LL(\theta)$ based on $f(X_i|\theta)$.
2. Differentiate $LL(\theta)$ and set to 0.
3. Solve simultaneous equations for $\theta_{MLE}$. 
MLE with Bernoulli

\[ \theta_{MLE} = \arg \max_\theta LL(\theta) = \arg \max_\theta \sum_{i=1}^n \log f(X_i|\theta) \]

Consider the sample of \( n \) IID RVs, \( X_1, X_2, \ldots, X_n \)

Assume model: \( X_i \sim \text{Ber}(p) \)

PMF \( f(X_i|\theta) = p^{x_i}(1-p)^{1-x_i} \), where \( x_i = 0 \) or \( x_i = 1 \)

What is \( \theta_{MLE} \)?

**Solution:**

1. \( LL(\theta) = \sum_{i=1}^n \log(p^{x_i}(1-p)^{1-x_i}) = \sum_{i=1}^n [X_i \log p + (1 - X_i) \log(1 - p)] \)
   
   \[ = Y \log p + (n - Y) \log(1 - p) \quad \text{where} \quad Y = \sum_{i=1}^n X_i \quad \text{(variable substitution)} \]

2. Differentiate w.r.t \( p \) and set to 0:
   
   \[ \frac{\partial LL(p)}{\partial p} = Y \frac{1}{p} + (n - Y) \frac{-1}{1 - p} = 0 \]

3. Solve.
   
   \[ p_{MLE} = \frac{Y}{n} = \frac{1}{n} \sum_{i=1}^n X_i \]
The issue with Maximum Likelihood Estimators

\[ \theta_{MLE} = \arg \max_{\theta} L(\theta) \]

The parameter \( \theta \) making observed sample the most likely is not always the “best” \( \theta \):

Consider a coin.

Experiment: Flip \( n = 5 \) times
Observe 5 heads

\[ p_{MLE} = \frac{\text{# of successes}}{n} = \frac{5}{5} = 1 \]

\[ L(p_{MLE}) = 1^5 = 1 \]

• This is very \textit{unbelievable}, because our \textit{prior knowledge} leads us to think that there is some possibility of flipping tails.

• But MLE says we can only use the data to estimate \( \theta \).

Solution: Maximum a Posteriori Estimators
Maximum A Posteriori Estimator

Maximum A Posteriori (MAP) Estimator:

$$\theta_{MAP} = \arg \max_\theta f(\theta|X_1, X_2, ..., X_n) = \arg \max_\theta \frac{g(\theta)}{\theta} \prod_{i=1}^{n} f(X_i|\theta)$$

If conjugate distribution known:
1. Assume imaginary trials
2. Update posterior with real experiment
3. Report mode of posterior

Laplace smoothing — Assume you have seen 1 imaginary trial per outcome.
Conjugate distributions

- When using MAP estimators, we have a choice of prior.
- But we need to know the distribution of our posterior in order to find its mode ($\theta_{MAP}$).

**Conjugate distributions:**
- For a given model (e.g., Bernoulli/Multinomial/Poisson)
- If we choose the conjugate distribution for the prior distribution of the parameter $\theta$ (e.g., Beta/Dirichlet/Gamma),
- We know the posterior distribution is the same (e.g., Beta/Dirichlet/Gamma),
- And therefore we can find the mode of the posterior.

Prior: $g(\theta)$
Likelihood: $\prod_{i=1}^{n} f(X_i|\theta)$
Posterior: $f(\theta|X_1, X_2, ..., X_n)$
$\theta_{MAP}$: $\arg \max_{\theta} f(\theta|X_1, X_2, ..., X_n)$
Interpreting the Multinomial MAP + Laplace:

Consider a 6-sided die.

Let $X \sim \text{Multi}(p_1, p_2, p_3, p_4, p_5, p_6)$

Experiment: Roll $n = 12$ times

Observe 3 ones, 2 twos, 0 threes, 3 fours, 1 five, 3 sixes

Recall MLE:

$$p_1 = \frac{3}{12}, \quad p_2 = \frac{2}{12}, \quad p_3 = \frac{0}{12}, \quad p_4 = \frac{3}{12}, \quad p_5 = \frac{1}{12}, \quad p_6 = \frac{3}{12}$$

What is Laplace estimator for $p_k$ for $k = 1, 2, \ldots, 6$?

Solution:

$$\hat{p}_1 = \frac{4}{18}, \quad \hat{p}_2 = \frac{3}{18}, \quad \hat{p}_3 = \frac{1}{18}, \quad \hat{p}_4 = \frac{4}{18}, \quad \hat{p}_5 = \frac{2}{18}, \quad \hat{p}_6 = \frac{4}{18}$$

No longer have 0 probability of rolling a 3!
MAP for Poisson

Conjugate distribution for Poisson is $\text{Gamma}(\alpha, \lambda)$.

Prior: $\text{Gamma}(\alpha, \lambda)$
Prior: saw $\alpha$ total imaginary events during $\lambda$ prior time periods

Experiment: Observe $n$ new events during next $k$ time periods

Posterior: $\text{Gamma}(\alpha + n, \lambda + k)$
(we will get this posterior,
for which the mode is known.)

MAP estimator: $\theta_{MAP} = \hat{\lambda} = \frac{\alpha + n}{\lambda + k}$
(for this model,)
(if we choose this prior,)
Consider a Poisson.

Imaginary experiment: Measure 5 time periods
Observe 10 events total

Real experiment: Measure 2 time periods
Observe 11 events.

What is MAP estimator for \( \hat{\lambda} \)?

Solution:

\[
\hat{\lambda} = \frac{10 + 11}{2 + 5} = \frac{21}{7} = 3
\]
Summary of MLE/MAP

Maximum Likelihood Estimator (MLE), $\theta_{MLE}$

Bernoulli

$$\hat{p} = \frac{\text{# of successes}}{n}$$

Binomial

$$\hat{p} = \frac{\text{# of successes}}{n}$$

Multinomial

$$\hat{p}_k = \frac{\text{(# times outcome k occurs)}}{n}$$

Maximum A Posteriori (MAP) with Laplace Smoothing, $\theta_{MAP}$

$$\hat{p} = \frac{(# \text{ of successes})+1}{n+2}$$

$$\hat{p} = \frac{(# \text{ of successes})+1}{n+2}$$

$$\hat{p}_k = \frac{(# \text{ times outcome k occurs})+1}{n+m}$$

Caveat for MLE/MAP: need independence of trials!
### Summary of MLE/MAP

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Maximum Likelihood Estimator (MLE), $\theta_{MLE}$</th>
<th>Maximum A Posteriori (MAP), $\theta_{MAP}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>$\hat{\lambda} = \frac{\text{# of successes}}{n}$</td>
<td>$\hat{\lambda} = \frac{(\text{# of successes}) + (\text{imag. successes})}{n + (\text{imag. intervals})}$</td>
</tr>
<tr>
<td>Normal</td>
<td>$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$, $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2$</td>
<td>(not discussed)</td>
</tr>
<tr>
<td>Uniform</td>
<td>$\hat{\alpha} = \min(X_1, X_2, ..., X_N)$, $\hat{\beta} = \max(X_1, X_2, ..., X_N)$</td>
<td>(not discussed)</td>
</tr>
</tbody>
</table>

Additional MLE caveat: suffers from small samples
Goals for today

Machine Learning
- Introduction
- Linear Regression
- Brute force Classification
- Naïve Bayes
- Evaluation tools: Precision, Recall

"Started from the bottom now we’re here" – Drake, 2013
Sampling use cases

Sample of population → Bootstrapping

Parameter Estimation

Estimates of $E[X]$ or $\text{Var}(X)$, confidence interval, p-values

Sample with features and labels → Machine Learning

Predict labels on new data

Model of $X$
What is Machine Learning?

Many different forms, but we focus on the problem of prediction. Sample data is now:

- Features $X = \langle X_1, X_2, \ldots, X_m \rangle$
- Label $Y$

Goal: Learn a function $\hat{Y} = g(X)$ to predict $Y$

1. **Training**: observe relationship from pre-existing data
   
   Known stock patterns:
   - $X$: economic indicators for stock
   - $Y$: stock’s price tomorrow

   Known medical records:
   - $X$: clinical/demographic data
   - $Y$: has cancer by age 40

2. **Prediction**:
   
   On a new, unseen stock:
   - $X$: economic indicators for this stock
   - $\hat{Y}$: stock’s price tomorrow?
     
   $Y$ continuous

   On a new person:
   - $X$: clinical/demographic data
   - $\hat{Y}$: cancer by age 40?
     
   $Y$ discrete

   Past purchases on credit card:
   - $X$: purchase details
   - $Y$: fraud or not

   A new purchase:
   - $X$: purchase details
   - $\hat{Y}$: fraud?
     
   $Y$ discrete
“And machine-learning algorithms developed to merge and rank large sets of Google search results allow us to combine hundreds of factors to classify spam.”

What is Machine Learning?

Many different forms, but we focus on the problem of prediction. Sample data is now:

- Features $X = \langle X_1, X_2, \ldots, X_m \rangle$
- Label $Y$

Goal: Learn a function $\hat{Y} = g(X)$ to predict $Y$

1. **Training**: observe relationship from pre-existing data

Sample size $N$:

$$
\begin{align*}
X^{(1)} &= \langle X_1 = 1, X_2 = 3, \ldots, X_m = 3.5 \rangle, & Y^{(1)} &= 1 \\
X^{(2)} &= \langle X_1 = 0, X_2 = 1, \ldots, X_m = -1.9 \rangle, & Y^{(2)} &= 0 \\
& \vdots \\
X^{(N)} &= \langle X_1 = 1, X_2 = 20, \ldots, X_m = 1.2 \rangle, & Y^{(N)} &= 0
\end{align*}
$$

- M-dimensional IID RV (feature)
- 1-D dependent (label)

2. **Prediction**: $Y$ discrete: prediction called “classification”

$Y$ continuous: prediction called “regression”
How do we find $g(X)$?

Mathematically:

1. Determine a parametric form for $\hat{Y} = g(X)$ with parameters $\theta$
   
   Train on data: $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)})$

2. Estimate the best parameters $\theta$ according to an objective function

   Y continuous: minimize mean squared error MSE “loss”, $E \left[ (Y - g(X))^2 \right]$

   Y discrete: maximize likelihood of $Y$ given features $X$, $g(X) = \arg\max_y \hat{P}(Y|X)$

Prediction on test data: Given a new input $X$, use $g(X)$ to predict the “class” $\hat{Y}$
Linear Regression (as a warmup)

Observe: Single variable $X$
Output: Real value $Y$ (continuous output)
Model: Assume linear: $\hat{Y} = g(X) = aX + b$
Training data: $N$ training pairs $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)})$
Goal: $\hat{Y} = aX + b$ where $a, b = \arg\min_{a,b} E \left[ (Y - g(X))^2 \right]$

1. $E \left[ (Y - g(X))^2 \right] = E \left[ (Y - (aX + b))^2 \right] = E[(Y - aX - b)^2]$

2. Compute derivatives w.r.t $a$ and $b$:

$$\frac{\partial}{\partial a} E[(Y - aX - b)^2] = E[-2X(Y - aX - b)] = -2E[XY] + 2aE[X^2] + 2bE[X]$$

$$\frac{\partial}{\partial b} E[(Y - aX - b)^2] = E[-2(Y - aX - b)] = -2E[Y] + 2aE[X] + 2b$$
Linear Regression

Observe: Single variable X
Output: Real value Y (continuous output)
Model: Assume linear: \( \hat{Y} = g(X) = aX + b \)
Training data: \( N \) training pairs \((x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(N)}, y^{(N)})\)
Goal: \( \hat{Y} = aX + b \) where \( a, b = \text{argmin}_{a,b} E \left[ (Y - g(X))^2 \right] \)

3. Set derivatives to 0 and solve simultaneous equations:
\[
a = -\frac{E[XY] - E[X]E[Y]}{E[X^2] - (E[X])^2} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} = \rho(X,Y) \frac{\sigma_Y}{\sigma_X} \quad b = E[Y] - aE[X] = \mu_y - \rho(X,Y) \frac{\sigma_Y}{\sigma_X} \mu_X
\]

4. Substitution:
\[g(X) = \rho(X,Y) \frac{\sigma_Y}{\sigma_X} (X - \mu_X) + \mu_Y\]

5. Estimate parameters based on observed training data:
\[
\hat{Y} = g(X = x) = \hat{\rho}(X,Y) \frac{\hat{\sigma}_Y}{\hat{\sigma}_X} (x - \bar{X}) + \bar{Y}
\]
Classification

Observe: Discrete observations $X = \langle X_1, X_2, \ldots, X_m \rangle$

Output: Discrete “class”

Model: $\hat{Y} = g(X) = \arg\max \hat{P}(Y|X)$

Note (via Bayes): $\hat{Y} = \arg\max \hat{P}(X, Y) = \arg\max \hat{P}(X|Y)\hat{P}(Y)$

Let $X_i, Y$ be indicator variables of liking a franchise.

$$[1, 1, 0] \xrightarrow{\hat{P}(X, Y)} \hat{P}(X, Y = 1) \approx 0.21$$

$$\xrightarrow{\text{argmax}_Y} \hat{Y} = 1$$

$$\hat{P}(X, Y = 0) \approx 0.14$$
Classification

Observe: Discrete variable $\mathbf{X} = \langle X_1, X_2, \ldots, X_m \rangle$

Output: Discrete output $Y$

Model: $\hat{Y} = g(\mathbf{X}) = \arg\max_y \hat{P}(Y|\mathbf{X}) = \arg\max_y \hat{P}(\mathbf{X}, Y)$

Training data: $N$ training pairs $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)})$

What is our best estimate of $\hat{P}(\mathbf{X}, Y)$?

- Recall: given a sample of size $n$, $\theta_{MLE}$ maximizes likelihood of seeing that particular sample. Each $X_i$ in sample is IID.

In our training data:

- Each $(x^{(i)}, y^{(i)})$ is a RV in our sample. $(x^{(i)}, y^{(i)})$’s are IID.
- $P(\mathbf{X}, Y)$ can be modeled as a multinomial distribution!
- # of outcomes: total # of probabilities in $P(\mathbf{X}, Y)$ space

What is $\theta_{MLE}$ that maximizes our likelihood of seeing the training data? What is $\theta_{MAP}$ that maximizes $\theta$ given the training data and an assumed prior?
Classification

Training data: $N$ training pairs $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)})$?

- What is $\theta_{MLE}$ that maximizes our likelihood of seeing the training data?

$$\theta_{MLE} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \prod_{i=1}^{n} f(x^{(i)}, y^{(i)} | \theta)$$

Model: $(X, Y)$ is multinomial

- What is $\theta_{MAP}$ that maximizes $\theta$ given the training data and a Laplace prior?

$$\theta_{MAP} = \arg\max_{\theta} g(\theta) \prod_{i=1}^{n} f(x^{(i)}, y^{(i)} | \theta)$$

Prior: Dirichlet with 1 imaginary outcome for each unique $(x^{(i)}, y^{(i)})$ in training data

Model: $(X, Y)$ is multinomial

Train:
1. MLE: Count # of times each pair $(x,y)$ appears in training data
2. MAP using Laplace prior: add 1 to all the MLE counts
3. Normalize to get true distribution (sum to 1)

Test/Predict:
- Given a new data point $X$, predict $\hat{Y} = \arg\max_{y} \hat{P}(X, Y)$
Training for cats

Observe: X in {1,2,3,4}
- X denotes location of photo: bedroom (1), bathroom (2), patio (3), fridge (4)

Output: Y from {0, 1}
- Y denotes if cat in photo (1) or not (0)

1. What are MLE estimates of $\hat{p}_{X,Y}(x, y)$?
2. What are Laplace estimates of $\hat{p}_{X,Y}(x, y)$?

### MLE estimate

<table>
<thead>
<tr>
<th>Y</th>
<th>X</th>
<th>p_y(y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>no cat</td>
<td>1</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.40</td>
</tr>
<tr>
<td>cat</td>
<td></td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.40</td>
</tr>
</tbody>
</table>

\[ \hat{p}_{MLE} = \frac{\text{count in cell}}{\text{total # data points}} \]

### Laplace (MAP) estimate

<table>
<thead>
<tr>
<th>Y</th>
<th>X</th>
<th>p_y(y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>no cat</td>
<td>1</td>
<td>0.069</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.138</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.190</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.362</td>
</tr>
<tr>
<td>cat</td>
<td></td>
<td>0.103</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.069</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.017</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.241</td>
</tr>
</tbody>
</table>

\[ \hat{p}_{\text{Laplace}} = \frac{\text{count in cell} + 1}{\text{total # data points} + \text{total # cells}} \]

\[ \hat{Y} = \text{argmax} \hat{P}(X, Y) \]
Testing for cats

Observe: \( X \) in \{1,2,3,4\}  
Location of photo: bedroom (1), bathroom (2), patio (3), fridge (4)
Output: \( Y \) from \{0, 1\}:  
Cat in photo (1) or not (0)

You are now shown a new photo of a fridge, so \( X = 4 \).

1. What is your prediction \( \hat{Y} \) for whether there is a cat (or not) in the photo?
2. Given \( X = 4 \), what is probability of having a cat in the photo?

### MLE estimate

<table>
<thead>
<tr>
<th>( Y )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>( p_Y(y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>no cat</td>
<td>0.06</td>
<td>0.14</td>
<td>0.20</td>
<td>0.40</td>
<td>0.80</td>
</tr>
<tr>
<td>cat</td>
<td>0.10</td>
<td>0.06</td>
<td>0.04</td>
<td>0.00</td>
<td>0.20</td>
</tr>
<tr>
<td>( p_X(x) )</td>
<td>0.16</td>
<td>0.20</td>
<td>0.24</td>
<td>0.40</td>
<td>1.00</td>
</tr>
</tbody>
</table>

### Laplace (MAP) estimate

<table>
<thead>
<tr>
<th>( Y )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>( p_Y(y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>no cat</td>
<td>0.069</td>
<td>0.138</td>
<td>0.190</td>
<td>0.362</td>
<td>0.759</td>
</tr>
<tr>
<td>cat</td>
<td>0.103</td>
<td>0.069</td>
<td>0.052</td>
<td>0.017</td>
<td>0.241</td>
</tr>
<tr>
<td>( p_X(x) )</td>
<td>0.172</td>
<td>0.207</td>
<td>0.242</td>
<td>0.379</td>
<td>1.00</td>
</tr>
</tbody>
</table>

1. Predict: \( \hat{Y} = 0 \) (no cat)
2. Absolutely, positively no chance of cat

1. Predict: \( \hat{Y} = 0 \) (no cat)
2. Small chance \( \rightarrow \) “never say never” 😐
The issue with multiple observables

Observe:  Discrete variable $\mathbf{X} = \langle X_1, X_2, \ldots, X_m \rangle$ each with 2 outcomes
Predict:  Discrete output $Y$ with 2 outcomes
Model:  $\hat{Y} = g(\mathbf{X}) = \operatorname*{argmax}_y \hat{P}(Y|\mathbf{X}) = \operatorname*{argmax}_y \hat{P}(\mathbf{X}, Y)$

How large is your lookup table $\hat{P}(\mathbf{X}, Y)$ if:

1. $m = 3$?
   
   $2^3 \cdot 2$

2. $m = 10$?
   
   $2^{10} \cdot 2$

3. $m = 10,000$?
   
   $2^{10,000} \cdot 2$

- Size of PMF table is \textbf{exponential} in $m!$ (e.g., $O(2^m)$)
- Need ridiculous amount of data for good probability estimates
- Likely to have many 0’s in table \textbf{(bad times)}
- Need to consider a simpler model.
Break

Attendance: tinyurl.com/cs109summer2018
Naïve Bayes Classifier

Observe: Discrete variable $X = <X_1, X_2, ..., X_m>$ each with 2 outcomes
Predict: Discrete output $Y$ with 2 outcomes

Objective: $\hat{Y} = g(X) = \arg\max_y \hat{P}(Y|X) = \arg\max_y \hat{P}(X,Y) = \arg\max_y \hat{P}(X|Y)\hat{P}(Y)$

Naïve Bayes assumption:
• Assume $X_1, X_2, ..., X_m$ conditionally independent given $Y$.
• You don’t really believe this...
• But it is an approximation we make to be able to make predictions

Therefore: $\hat{Y} = \arg\max_y \hat{P}(X|Y)\hat{P}(Y) = \arg\max_y \prod_{i=1}^m \hat{P}(X_i|Y) \hat{P}(Y)$
• Computation of PMF table is now linear in $m$: $O(m)$
• Don’t need as much data to get good probability estimates
Naïve Bayes Classification

Objective: $\hat{Y} = g(X) = \arg\max_y \hat{P}(Y|X)$

$$= \arg\max_y \hat{P}(X,Y) = \arg\max_y \prod_{i=1}^{m} \hat{P}(X_i|Y) \hat{P}(Y)$$

Train
1. Compute $\hat{P}(X_i|Y = y)$ $i = 1, \ldots, m$:
   a. *Bernoulli* MLE (# times $(X_i = 0, Y = y)$ vs $(X_i = 1, Y = y)$)
   b. Laplace estimate: +1 to MLE estimates
   c. Normalize to get true distribution (sum to 1)
2. Compute $\hat{P}(Y)$: (Note: no Laplace smoothing for class probs.)
   a. *Bernoulli* MLE (# times $Y = 0$ vs $Y = 1$)
   b. Normalize by # samples to get true distribution (sum to 1)

Classify new $X = \langle X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m \rangle$
1. Compute $\hat{P}(X,Y = y) = \prod_{i=1}^{m} \hat{P}(X_i = x_i|Y = y) \hat{P}(Y = y)$
   for $Y = 0$ and $Y = 1$
2. Return the label (“class”) that led to higher probability
Naïve Bayes + MLE for movies

Observe: indicator variables $X_1, X_2$
- $X_1$: “likes Star Wars”
- $X_2$: “likes Harry Potter”

Output: $Y$ indicator variable:
- $Y$: “likes Pokemon”

- $X_1 = 1$
  Likes Star Wars
- $X_2 = 1$
  Likes Harry Potter
- $Y = 1$
  Likes Pokemon
Naïve Bayes + MLE for movies

Observe: indicator variables $X_1, X_2$
- $X_1$: “likes Star Wars”
- $X_2$: “likes Harry Potter”

Output: $Y$ indicator variable:
- $Y$: “likes Pokemon”

Train:
1. What are MLE estimates of $\hat{p}_{X_i|Y}(x_i|y)$?
2. What are MLE estimates of $\hat{p}_Y(y)$?

Classify/Predict/Test:
3. New person: likes Star Wars ($X_1 = 1$) but not Harry Potter ($X_2 = 0$). Will they like Pokemon?

Train data: $n=30$ points

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$Y$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$Y$</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$X_2$</th>
<th>$Y$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$Y$</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>7</td>
<td>10</td>
</tr>
</tbody>
</table>

$\hat{p}_{X_1|Y}(x_1|y) = \frac{\#(X_1 = x_1, Y = y)}{\#(Y = y)}$

$\hat{p}_{X_2|Y}(x_2|y) = \frac{\#(X_2 = x_2, Y = y)}{\#(Y = y)}$

<table>
<thead>
<tr>
<th>$Y$</th>
<th>$\hat{p}_Y(y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.43</td>
</tr>
<tr>
<td>1</td>
<td>0.57</td>
</tr>
</tbody>
</table>

$\hat{p}_Y(y) = \frac{\#(Y = y)}{n}$
Naïve Bayes + MLE for movies

Observe: indicator variables $X_1, X_2$
- $X_1$: “likes Star Wars”
- $X_2$: “likes Harry Potter”

Output: $Y$ indicator variable:
- $Y$: “likes Pokemon”

Train:
1. What are MLE estimates of $\hat{p}_{X_i|Y}(x_i|y)$?
2. What are MLE estimates of $\hat{p}_Y(y)$?

Classify/Predict/Test:
3. New person: likes Star Wars ($X_1 = 1$) but not Harry Potter ($X_2 = 0$). Will they like Pokemon?

| $Y$ | $X_1$ | $X_2$ | $\hat{p}_{X_1|Y}(x_1|y)$ | $\hat{p}_{X_2|Y}(x_2|y)$ | $\hat{p}_Y(y)$ |
|-----|-------|-------|--------------------------|--------------------------|---------------|
| 0   | 0     | 0     | 0.23                     | 0.38                     | 0.43          |
|     | 1     | 0     | 0.24                     | 0.41                     | 0.43          |
| 1   | 0     | 1     | 0.77                     | 0.62                     | 0.57          |
|     | 1     | 1     | 0.76                     | 0.59                     | 0.57          |

$\hat{P}(X,Y = 0) = \hat{P}(X_1 = 1|Y = 0) \hat{P}(X_2 = 0|Y = 0) \hat{P}(Y = 0) = \hat{p}_{X_1|Y}(1|0)\hat{p}_{X_2|Y}(0|0)\hat{p}_Y(0) = 0.77 \cdot 0.38 \cdot 0.43 \approx 0.13$

$\hat{P}(X,Y = 1) = \hat{P}(X_1 = 1|Y = 1) \hat{P}(X_2 = 0|Y = 1) \hat{P}(Y = 1) = \hat{p}_{X_1|Y}(1|1)\hat{p}_{X_2|Y}(0|1)\hat{p}_Y(1) = 0.76 \cdot 0.41 \cdot 0.57 \approx 0.18$

$\hat{Y} = 1$: predict “like Pokemon”
Naïve Bayes + Laplace for movies

Observe: indicator variables $X_1, X_2$
- $X_1$: “likes Star Wars”
- $X_2$: “likes Harry Potter”

Output: $Y$ indicator variable:
- $Y$: “likes Pokemon”

Train:
1. What are Laplace estimates of $\hat{p}_{X_1|Y}(x_i|y)$?
2. What are MLE estimates of $\hat{p}_Y(y)$?

Train data: $n=30$ points

\[
\begin{array}{c|cc}
Y & 0 & 1 \\
\hline
0 & 3 & 10 \\
1 & 4 & 13 \\
\end{array}
\]

\[
\begin{array}{c|cc}
Y & 0 & 1 \\
\hline
0 & 0.27 & 0.73 \\
1 & 0.26 & 0.74 \\
\end{array}
\]

\[
\begin{array}{c|cc}
Y & 0 & 1 \\
\hline
0 & 0.40 & 0.60 \\
1 & 0.42 & 0.58 \\
\end{array}
\]
Summary so far

Observe: Discrete variable $X = < X_1, X_2, ..., X_m >$ each with 2 outcomes
Output: Discrete output $Y$ with 2 outcomes

Classification:

$$\hat{Y} = g(X) = \arg\max_y \hat{P}(Y|X) = \arg\max_y \hat{P}(X,Y) = \arg\max_y \hat{P}(X|Y)\hat{P}(Y)$$

(by definition) (equivalent) (equivalent)

Naïve Bayes assumption:

• Assume $X_i$’s conditionally independent given $Y$
• Reduce the probabilities we need to compute from $O(2^m)$ to $O(m)$
• Training: Use MLE or Laplace estimate for $\hat{P}(X_i|Y)$, MLE estimate for $\hat{P}(Y)$
• Testing: Compute:

$$\hat{Y} = \arg\max_y \hat{P}(Y|X) = \arg\max_y \prod_{i=1}^{m} \hat{P}(X_i = x_i|Y = y) \hat{P}(Y = y)$$
Log-probabilities

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

What if \( m \) is large and \( \hat{P}(X_i|Y) \) is small?

Solution: Take logs:

\[ \hat{Y} = \arg \max_{y} \log \hat{P}(Y|X) = \arg \max_{y} \log \left( \hat{P}(Y = y) \prod_{i=1}^{n} \hat{P}(X_i = x_i|Y = y) \right) \]

\[ = \arg \max_{y} \left( \log \hat{P}(Y = y) + \sum_{i=1}^{n} \log \left( \hat{P}(X_i = x_i|Y = y) \right) \right) \]

Works well for high-dimensional data
Bags of words as a multinomial

Example document:
“Pay for Viagra with a credit card. Viagra is great. So are credit-cards. Risk-free Viagra. Click for free.”

\[ n = 18 \]

\[
P( \text{this document} | \text{spam writer}) = \frac{n!}{2!2! \ldots 2!} p_{\text{viagra}}^2 p_{\text{free}}^2 \cdots p_{\text{for}}^2
\]

Note \( P( \text{viagra} | \text{spam writer}) \gg P( \text{viagra} | \text{writer = you}) \)
Bag of Words with Naïve Bayes

\[
\hat{Y} = \arg\max_y \left( \log \hat{P}(Y = y) + \sum_{i=1}^{n} \log \left( \hat{P}(X_i = x_i|Y = y) \right) \right)
\]

Goal: Predict if an email is spam or not.

Known: Lexicon of \( m \) words (e.g., English language \( m \approx 100,000 \))

Observe: \( m \) indicator variables \( X = \langle X_1, X_2, ..., X_m \rangle \) (no longer counts)

\( X_i \) denotes if word \( i \) appeared in an email or not

Output: \( Y \) indicator variable denoting if email is spam or not

Preprocess: \( N \) previous emails with the following info per email:

- \( X = \langle X_1, X_2, ..., X_m \rangle \) noting for each word if it appeared in this email
- \( Y \) label marking this email as spam or not
Spam Classifier

Training

1. Go through all training emails and update all counts:
   \[(X_i = 1, Y = \text{spam}), \ (X_i = 1, Y = \text{not spam})\]
   \[(X_i = 0, Y = \text{spam}), \ (X_i = 0, Y = \text{not spam})\]
   
   Sanity check: \(#(X_i = 1, Y = \text{spam}) + #(X_i = 0, Y = \text{spam}) = #(Y = \text{spam})\)

2. Since many words (in the lexicon of 100,000 words) are likely to not appear at all in the training emails,
   
   Use Laplace estimate:
   \[
   \hat{P}(X_1 = 1|Y = \text{spam}) = \frac{\#(X_1 = x_1, Y = y) + 1}{\#(Y = y) + 2}
   \] (and so on)

Classifying

• Classify new email:
  \[
  \hat{Y} = \arg\max_y \hat{P}(Y|X) = \arg\max_y \hat{P}(X|Y)\hat{P}(Y)
  \]

• Employ Naïve Bayes assumption:
  \[
  \hat{Y} = \arg\max_y \left( \log \hat{P}(Y = y) + \sum_{i=1}^{n} \log \left( \hat{P}(X_i = x_i|Y = y) \right) \right)
  \]
Checking performance

After training, we can test with another set of data.

Test data: also has features $X = \langle X_1, X_2, \ldots, X_m \rangle$ and known $Y$ values

Can see how often we were right/wrong in predictions for $Y$.

Metrics:

- **Precision**: $(\# \text{ correctly predicted class } Y)/(\# \text{ predicted class } Y)$
- **Recall**: $(\# \text{ correctly predicted class } Y)/(\# \text{ real class } Y \text{ messages})$
- **Accuracy**: $(\# \text{ correct across all classes})/(\text{total test emails})$

Spam data:

- Email dataset: 1789 emails
  (1578 spam, 211 non-spam)
- Randomly choose 1538 emails for training
- Remaining 251 emails for testing learned classifier

<table>
<thead>
<tr>
<th></th>
<th>Spam Precision</th>
<th>Spam Recall</th>
<th>Non-spam Precision</th>
<th>Non-spam Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words only</td>
<td>97.1%</td>
<td>94.3%</td>
<td>87.7%</td>
<td>93.4%</td>
</tr>
</tbody>
</table>
Congratulations!

You’ve just learned Machine Learning in a nutshell.

1. Have some training data.

2. Assume some relationship between the features and labels.

3. Get estimates of relationship from training data.

4. Use estimates to predict on test data using assumptions.

41

\[ \hat{Y} = \arg\max_y \left( \log \hat{P}(Y = y) + \sum_{i=1}^{n} \log \left( \hat{P}(X_i = x_i | Y = y) \right) \right) \]