PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Introduction to Machine Learning

MLEARN 510A – Lesson 9

Recap of Lesson 8

- Shrinkage Methods
- Polynomial Regression
- Step Functions
- Basis Functions
- Regression Splines
- Local Regression
- Generalized Additive Models

Course Outline

1. Introduction to Statistical Learning
2. Linear Regression
3. Classification
4. Model Building, Part 1
5. Model Building, Part 2
6. Resampling Methods
7. Linear Model Selection and Regularization
8. Moving Beyond Linearity
9. **Bayesian Analysis**
10. Dimensionality Reduction

Outline of Lesson 9

- Review of Probability
- Conditional Probability
- Bayes Theorem
- Application of Bayes Theorem
- Bayesian Networks
- Reasoning with BN
- Naïve Bayes

Probability: Intuition

- Probability theory is used to model systems where the outcomes are either inherently random, or simply too complex to be completely known
- For e.g., result from rolling a fair die is uncertain
- Probability of an event encodes the fraction of the times that outcome would occur with repeated experiments

Probability: Intuition

- Outcome: Result/realization of an experiment
- Sample space: the set of all possible outcomes of the experiment
- Event: a subset of the sample space
- Discrete random variable: a random variable which can only take a countable number of values
- Probability: this is the fraction of the times that you see the event occurring

Probability: Example

- Suppose I flip three coins, what is the probability I get exactly two heads?

$$P(A) = \frac{\text{No. of outcomes favourable to the occurrence of } A}{\text{Total number of equally likely outcomes}} = \frac{n(A)}{n(S)}$$

Probability Distribution

- A probability distribution is a function that links each outcome of a statistical experiment with its probability of occurrence

- Probability distribution from a fair die

i	$\dots$	1	2	3	4	5	6	$\dots$
$\mathbb{P}\{i\}$	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0

- Probability distribution from an unfair die

i	$\dots$	1	2	3	4	5	6	$\dots$
$\mathbb{P}\{i\}$	0	0.1	0.2	0.1	0.3	0.2	0.1	0

Conditional Probability

- Conditional probability is a measure of the probability of occurrence of an event given that another event has occurred

- Motivation: Partial information

$$\mathbb{P}(\mathrm{A}|\mathrm{B}) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

- Fundamental importance in probability theory

Conditional Probability: Examples

When rolling a fair die what is:

- $\mathbb{P}\{X\ odd\}$
- $\mathbb{P}\{X\ odd \mid X \geq 4\}$
- $\mathbb{P}\{X\ odd \mid X \leq 3\}$
- $\mathbb{P}\{X\ odd \mid X \geq 3\}$

Bayes Rule

- Conditional probability $\mathbb{P}(\mathrm{A}|\mathrm{B}) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ can be expressed as

$\mathbb{P}(A \cap B) = \mathbb{P}(\mathrm{A}|\mathrm{B}) \times \mathbb{P}(B)$

- Conditional probability $\mathbb{P}(\mathrm{B}|\mathrm{A}) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}$ can be expressed as

$\mathbb{P}(B \cap A) = \mathbb{P}(\mathrm{B}|\mathrm{A}) \times \mathbb{P}(A)$

- This implies $\mathbb{P}(\mathrm{A}|\mathrm{B}) \times \mathbb{P}(B) = \mathbb{P}(\mathrm{B}|\mathrm{A}) \times \mathbb{P}(A)$
- Dividing both side by $\mathbb{P}(B)$ we obtain

Bayes Rule

$$\mathbb{P}(\mathrm{A}|\mathrm{B}) = \frac{\mathbb{P}(\mathrm{B}|\mathrm{A}) \times \mathbb{P}(A)}{\mathbb{P}(B)}$$

Bayes Inference

- Bayesian inference is the process of confronting alternative hypotheses with new data and using Bayes' Theorem to update your beliefs in each hypothesis

- Bayes' theorem: $\mathbb{P}(\mathrm{A}|\mathrm{B}) = \frac{\mathbb{P}(B|A)\times\mathbb{P}(A)}{\mathbb{P}(B)}$

- To use Bayes' theorem for scientific inference, it's essential to replace the marginal denominator $\mathbb{P}(B)$ as the sum of the joint probabilities that make it up:

$$\mathbb{P}(\mathrm{A}|\mathrm{B}) = \frac{\mathbb{P}(B|A) \times \mathbb{P}(A)}{\mathbb{P}(A \cap B) + \mathbb{P}({\sim}A \cap B)}$$

Bayes Inference

	A	~A	Marginal
B	$\mathbb{P}(A \cap B)$	$\mathbb{P}(\sim A \cap B)$	$\mathbb{P}(B)$
~B	$\mathbb{P}(A \cap \sim B)$	$\mathbb{P}(\sim A \cap \sim B)$	$\mathbb{P}(\sim B)$
Marginal	$\mathbb{P}(A)$	$\mathbb{P}(\sim A)$	Total: 1.0

Quiz: The Seattle Rain Problem

$$\mathbb{P}(\mathrm{A}|\mathrm{B}) = \frac{\mathbb{P}(B|A) \times \mathbb{P}(A)}{\mathbb{P}(A \cap B) + \mathbb{P}(\sim A \cap B)}$$

- *"You're about to get on a plane to Seattle. You want to know if you should bring an umbrella. You call 3 random friends of yours who live there and ask each independently if it's raining. Each of your friends has a 2/3 chance of telling you the truth and a 1/3 chance of messing with you by lying. All 3 friends tell you that "Yes" it is raining. What is the probability that it's actually raining in Seattle?*

Quiz: Bayesian Inference

- One percent of women at age forty who participate in routine screening have breast cancer; 80% of women with breast cancer will have a positive mammogram (test), while 9.5% of women without breast cancer will also get a positive result. A woman in this age group had a positive mammogram in a routine screening. What is the probability that she actually has breast cancer?

Applications to Models and Data

- Suppose we have $H_1, \ldots, H_k$ competing hypotheses, and we observe some data D that helps us decide

$$\mathbb{P}\{H_i|D\} = \frac{\mathbb{P}\{D|H_i\}.\,\mathbb{P}\{Hi\}}{\sum_{j=1}^{n} \mathbb{P}\{D|Hj\}\,.\,P\{Hj\}}$$

- Start with a problem (scientific question)
- Set forth two or more alternative hypotheses
- Assign a prior probability that each alternative hypothesis is true
- Next, collect data
- Use Bayes' theorem to update the probability for each hypothesis considered

Example

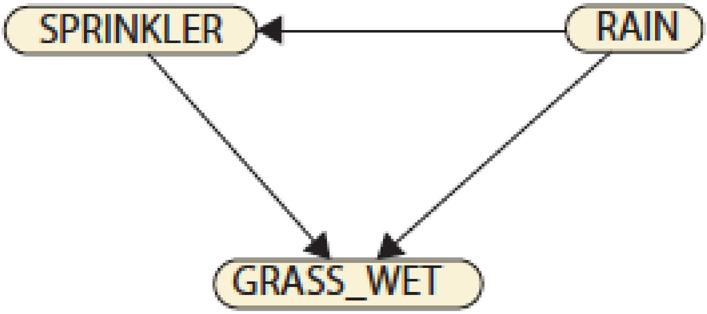
- Suppose I have a bag with 2 dice. The red one is fair, the blue one is only ones. I draw a random die, roll it, and get a one. What is the probability I drew the fair die?

- We can rephrase that example into the machine learning scheme:

Given the observations (i.e. the training data), predict whether the die is fair or unfair (i.e. binary classification)

Bayesian Networks

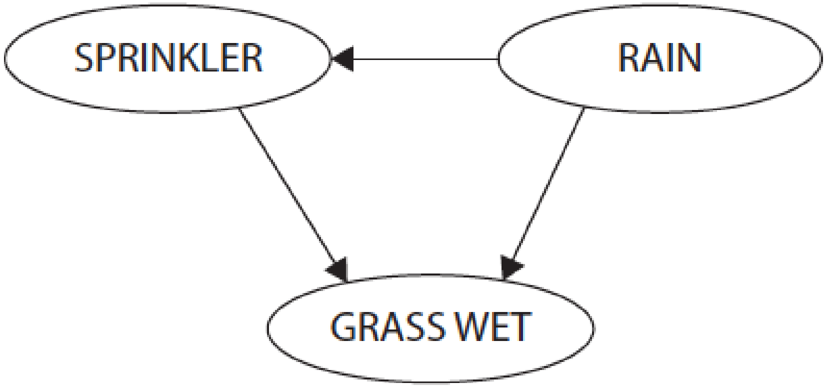
- A Bayesian Network is a probabilistic graphical model for depicting probabilistic relationships among a set of variables and their conditional dependencies via a directed acyclic graph (DAG)



- The direction of the link arrows roughly corresponds to causality

Conditional Probability Table

RAIN	SPRINKLER T	SPRINKLER F
F	0.4	0.6
T	0.01	0.99



RAIN T	RAIN F
0.2	0.8

SPRINKLER	RAIN	GRASS WET T	GRASS WET F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Conditional Probability Table

- A Bayesian network requires tables that hold conditional probabilities
- Nodes with no arrows leading to them have tables that provide **marginal probabilities**
- Nodes with arrows leading to them have tables that provide **conditional probabilities**

Directed Acyclic Graph

- A directed acyclic graph (DAG) is a graph that is directed and without cycles connecting the other edges

Reasoning With Bayesian Networks

- We can use the network to answer all kind of questions
  - If the grass is wet, what are the chances it was caused by rain?
  - If the chance of rain increases, how does that affect the amount of time I'll need to spend watering the lawn?

- To use the network, the relevant underlying conjoint tables must be computed

			GRASS WET	
SPRINKLER		RAIN	T	F
1	False	False		
2	False	True		
3	True	False		
4	True	True		

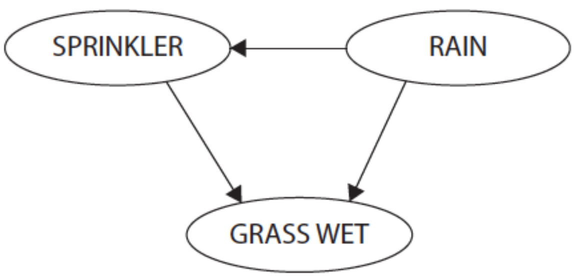
Reasoning With Bayesian Networks

- We compute each joint probability using the chain rule in probability

$$\mathbb{p}(A_4 \cap A_3 \cap A_2 \cap A_1) = \mathbb{p}(A_4|A_3 \cap A_2 \cap A_1) \times \mathbb{p}(A_3|A_2 \cap A_1) \times \mathbb{p}(A_2|A_1) \times \mathbb{p}(A_1)$$

- We can use the chain rule to compute the joint probability that the grass is wet, the sprinkler is on and it's raining

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Reasoning With Bayesian Networks

- Compute the joint probabilities using the chain rule

			GRASS WET		
SPRINKLER		RAIN	T	F	Sum
1	False	False	0.00000	0.48000	0.480
2	False	True	0.15840	0.03960	0.198
3	True	False	0.28800	0.03200	0.320
4	True	True	0.00198	0.00002	0.002
Sum →			**0.44838**	**0.55162**	**1.000**

- Bayesian network simplifies the number of calculations dramatically by computing only the required joint probabilities

- The network size grows linearly, with each new node doubling the number of parameters to estimate

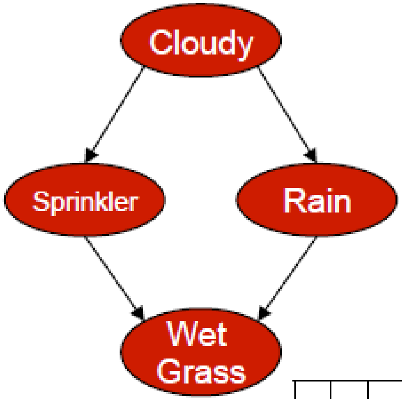
Quiz

- Given the Bayesian network

- What is the joint probability distribution P(C, S, R, W)?

> It is cloudy, what's the probability that the grass
> is wet?

+c	0.5
-c	0.5



+c	+s	.1
+c	-s	.9
-c	+s	.5
-c	-s	.5

+c	+r	.8
+c	-r	.2
-c	+r	.2
-c	-r	.8

+s	+r	+w	.99
+s	+r	-w	.01
+s	-r	+w	.90
+s	-r	-w	.10
-s	+r	+w	.90
-s	+r	-w	.10
-s	-r	+w	0
-s	-r	-w	1.0

Reasoning With Bayesian Networks

- Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG)

- Each node in the graph is a variable that has alternative states

- Each state occurs with some probability

- The nodes are linked with arrows, and the direction of the link roughly corresponds to causality.

- Bayes' theorem is at the heart of these connections

Bayesian Parameter Estimation

- Assume that a set of probability distribution parameters, θ, best explains the dataset *D*

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

$$posterior = \frac{likelihood * prior}{evidence}$$

- MLE maximized the likelihood function *p(D|θ)*

- MLE does not allow us to inject any prior belief *p(θ)* about any likely values of θ

- This is the **frequentist approach**

Bayesian Parameter Estimation

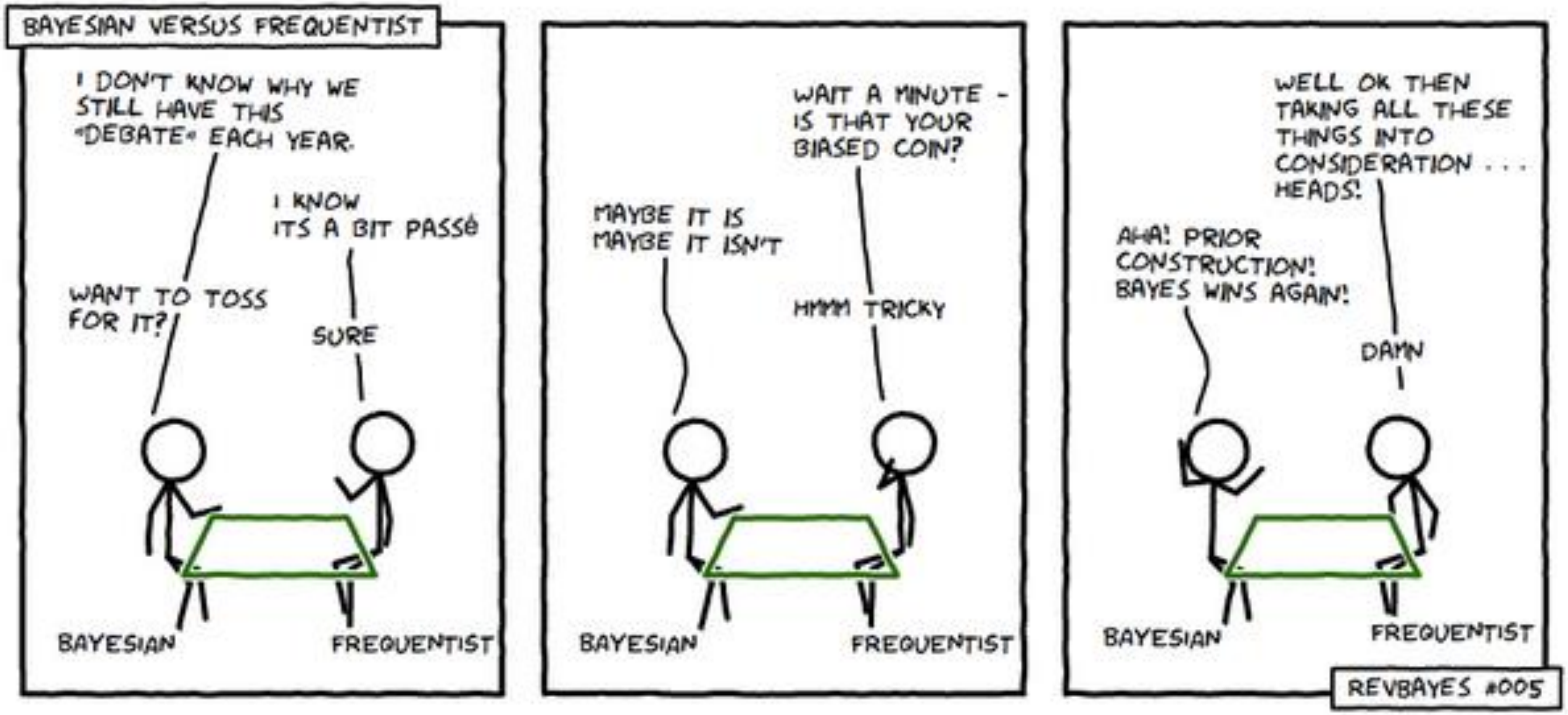
- What if we have reason to believe θ takes on a certain distribution?

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

$$posterior = \frac{likelihood * prior}{evidence}$$

- **Bayesian Estimation** treats θ as a random variable
- Fully calculates posterior *p(θ|D)*
- Maximize posterior distribution to get *Maximum a-posteriori Estimate (MAP)*

Bayesian Estimation vs. MLE



Naïve Bayes

- Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem

- Make "naive" assumption of conditional independence between every pair of features given the value of the class variable

- Bayes' theorem states the following relationship, given class variable y and dependent feature vector x_1 through x_n, :

$$P(y \mid x_1, \dots, x_n) = \frac{P(y) P(x_1, \dots, x_n \mid y)}{P(x_1, \dots, x_n)}$$

Naïve Bayes – Conditional Independence

- Using the naive conditional independence assumption

$$P(x_i|y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i|y),$$

- This relationship simplifies to

$$P(y \mid x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \dots, x_n)}$$

- Since $P(x_1,\dots,x_n)$ is constant given the input, we can use the following classification rule:

$$P(y \mid x_1, \dots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

Naïve Bayes – Pros and Cons

- Surprisingly powerful in some scenarios for e.g., document classification and spam filtering
- Can be extremely fast compared to more sophisticated methods
- Each distribution can be independently estimated as a one-dimensional distribution
- Performs poorly if features are not independent given class
- Feature independence assumption leads to poor estimation of class probabilities

Different Flavors of Naïve Bayes

- Several different types of Naïve Bayes depending on how we define $P(x_i|y)$
- Gaussian Naïve Bayes
- Multinomial Naïve Bayes
- Bernoulli Naïve Bayes

Gaussian Naïve Bayes

- Used when dealing with continuous data

- The likelihood of the features is assumed to be Gaussian

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

- The parameters σ_y and μ_y are estimated using maximum likelihood

```
>>> from sklearn.datasets import load_iris
>>> from sklearn.model_selection import train_test_split
>>> from sklearn.naive_bayes import GaussianNB
>>> X, y = load_iris(return_X_y=True)
>>> X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.5, random_state=0)
>>> gnb = GaussianNB()
>>> y_pred = gnb.fit(X_train, y_train).predict(X_test)
>>> print("Number of mislabeled points out of a total %d points : %d"
...       % (X_test.shape[0], (y_test != y_pred).sum()))
Number of mislabeled points out of a total 75 points : 4
```

Multinomial Naïve Bayes

- Used when dealing with discrete data

- Features represent the frequencies with which certain events have been generated by a multinomial distribution

- Feature vector **x** = [x_1, x_2, …, x_n] is a histogram with x_i counting the number of times event *i* was observed in a particular instance

- Typically used for document classification, with events representing the occurrence of a word in a single document

$$p(\mathbf{x} \mid C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i {p_{ki}}^{x_i}$$

Bernoulli Naïve Bayes

- Features are independent binary variables describing inputs
- Binary term occurrence features are used rather than term frequencies
- Elements x_i represent the presence or absence of the i^{th} term in the vocabulary
- Especially popular for classifying short text

$$p(\mathbf{x} \mid C_k) = \prod_{i=1}^{n} p_{ki}^{x_i} (1 - p_{ki})^{(1-x_i)}$$

Jupyter Notebook

- *Case Study*

ON-BRAND STATEMENT

FOR GENERAL USE

> What defines the students and faculty of the University of Washington? Above all, it’s our belief in possibility and our unshakable optimism. It’s a connection to others, both near and far. It’s a hunger that pushes us to tackle challenges and pursue progress. It’s the conviction that together we can create a world of good. And it’s our determination to Be Boundless. Join the journey at **uw.edu**.