Lecture 6: Probability Review

Aykut Erdem
October 2017
Hacettepe University
Last time… Regularization, Cross-Validation

\[ \tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2 \]

\[ \|\mathbf{w}\|^2 \equiv \mathbf{w}^T \mathbf{w} = w_0^2 + w_1^2 + \ldots + w_M^2 \]

<table>
<thead>
<tr>
<th>\ln \lambda = -\infty</th>
<th>\ln \lambda = -18</th>
<th>\ln \lambda = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w_0 )</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td>( w_1 )</td>
<td>232.37</td>
<td>4.74</td>
</tr>
<tr>
<td>( w_2 )</td>
<td>-5321.83</td>
<td>-0.77</td>
</tr>
<tr>
<td>( w_3 )</td>
<td>48568.31</td>
<td>-31.97</td>
</tr>
<tr>
<td>( w_4 )</td>
<td>-231639.30</td>
<td>-3.89</td>
</tr>
<tr>
<td>( w_5 )</td>
<td>640042.26</td>
<td>55.28</td>
</tr>
<tr>
<td>( w_6 )</td>
<td>-1061800.52</td>
<td>41.32</td>
</tr>
<tr>
<td>( w_7 )</td>
<td>1042400.18</td>
<td>-45.95</td>
</tr>
<tr>
<td>( w_8 )</td>
<td>-557682.99</td>
<td>-91.53</td>
</tr>
<tr>
<td>( w_9 )</td>
<td>125201.43</td>
<td>72.68</td>
</tr>
</tbody>
</table>

Figure credit: Fei-Fei Li, Andrej Karpathy, Justin Johnson
Today

• Probability Review
Basic Probability Review
Probability

• A is non-deterministic event
  – Can think of A as a boolean-valued variable

• Examples
  – A = your next patient has cancer
  – A = Rafael Nada wins French Open 2015
If I flip this coin, the probability that it will come up heads is 0.5

- **Frequentist Interpretation**: If we flip this coin many times, it will come up heads about half the time. *Probabilities are the expected frequencies of events over repeated trials.*

- **Bayesian Interpretation**: I believe that my next toss of this coin is equally likely to come up heads or tails. *Probabilities quantify subjective beliefs about single events.*

- Viewpoints play complementary roles in **machine learning**:
  - Bayesian view used to build models based on domain knowledge, and automatically derive learning algorithms
  - Frequentist view used to analyze worst case behavior of learning algorithms, in limit of large datasets

- From either view, basic mathematics is the same!
The Axioms Of Probability
Axioms of Probability

• $0 \leq P(A) \leq 1$
• $P(\text{empty-set}) = 0$
• $P(\text{everything}) = 1$
• $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
Interpreting the Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{empty-set}) = 0$
- $P(\text{everything}) = 1$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
Interpreting the Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{empty-set}) = 0$
- $P(\text{everything}) = 1$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

The area of A can’t get any smaller than 0

And a zero area would mean no world could ever have A true
Interpreting the Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{empty-set}) = 0$
- $P(\text{everything}) = 1$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

The area of $A$ can't get any bigger than 1

And an area of 1 would mean all worlds will have $A$ true
Interpreting the Axioms

- $0 \leq P(A) \leq 1$
- $P(\emptyset) = 0$
- $P(\text{everything}) = 1$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
Discrete Random Variables

\[ X \rightarrow \text{discrete random variable} \]

\[ \mathcal{X} \rightarrow \text{sample space of possible outcomes, which may be finite or countably infinite} \]

\[ x \in \mathcal{X} \rightarrow \text{outcome of sample of discrete random variable} \]
Discrete Random Variables

$X \rightarrow$ discrete random variable

$\mathcal{X} \rightarrow$ sample space of possible outcomes, which may be finite or countably infinite

$x \in \mathcal{X} \rightarrow$ outcome of sample of discrete random variable

$p(X = x) \rightarrow$ probability distribution (probability mass function)

$p(x) \rightarrow$ shorthand used when no ambiguity

$0 \leq p(x) \leq 1$ for all $x \in \mathcal{X}$

$$\sum_{x \in \mathcal{X}} p(x) = 1$$

$\mathcal{X} = \{1, 2, 3, 4\}$

uniform distribution

degenerate distribution
Joint Distribution
Marginalization

• Marginalization
  - Events: \( P(A) = P(A \text{ and } B) + P(A \text{ and not } B) \)
  
  - Random variables \( P(X = x) = \sum_{y} P(X = x, Y = y) \)
Marginal Distributions

\[ p(x, y) = \sum_{z \in \mathcal{Z}} p(x, y, z) \]

\[ p(x) = \sum_{y \in \mathcal{Y}} p(x, y) \]
Conditional Probabilities

- $P(Y=y \mid X=x)$
- What do you believe about $Y=y$, if I tell you $X=x$?
- $P($Rafael Nadal wins French Open 2015$)$?
- What if I tell you:
  - He has won the French Open 9/10 he has played there
  - Novak Djokovic is ranked 1; just won Australian Open
  - I offered a similar analysis last year and Nadal won
Conditional Probabilities

• $P(A \mid B) = \text{In worlds that where B is true, fraction where A is true}$

• Example
  - $H$: “Have a headache”
  - $F$: “Coming down with Flu”

\[ P(H) = \frac{1}{10} \]
\[ P(F) = \frac{1}{40} \]
\[ P(H \mid F) = \frac{1}{2} \]

Headaches are rare and flu is rarer, but if you’re coming down with flu there’s a 50-50 chance you’ll have a headache.
Conditional Distributions

\[ p(x, y \mid Z = z) = \frac{p(x, y, z)}{p(z)} \]
Independent Random Variables

Independent Random Variables

$$p(x, y) = p(x)p(y)$$
for all $$x \in \mathcal{X}, y \in \mathcal{Y}$$

Equivalent conditions on conditional probabilities:

$$p(x \mid Y = y) = p(x) \text{ and } p(y) > 0 \text{ for all } y \in \mathcal{Y}$$

$$p(y \mid X = x) = p(y) \text{ and } p(x) > 0 \text{ for all } x \in \mathcal{X}$$
Bayes Rule (Bayes Theorem)

\[ p(x, y) = p(x)p(y \mid x) = p(y)p(x \mid y) \]

\[ p(y \mid x) = \frac{p(x, y)}{p(x)} = \frac{p(x \mid y)p(y)}{\sum_{y' \in Y} p(y')p(x \mid y')} \]

\[ \propto p(x \mid y)p(y) \]

- A basic identity from the definition of conditional probability
- Used in ways that have no thing to do with Bayesian statistics!
- Typical application to learning and data analysis:

\[ Y \quad \longrightarrow \quad \text{unknown parameters we would like to infer} \]

\[ X = x \quad \longrightarrow \quad \text{observed data available for learning} \]

\[ p(y) \quad \longrightarrow \quad \text{prior distribution (domain knowledge)} \]

\[ p(x \mid y) \quad \longrightarrow \quad \text{likelihood function (measurement model)} \]

\[ p(y \mid x) \quad \longrightarrow \quad \text{posterior distribution (learned information)} \]
Binary Random Variables

- **Bernoulli Distribution:** Single toss of a (possibly biased) coin
  \[ \mathcal{X} = \{0, 1\} \]
  \[ 0 \leq \theta \leq 1 \]
  \[ \text{Ber}(x \mid \theta) = \theta \delta(x,1) (1 - \theta) \delta(x,0) \]

- **Binomial Distribution:** Toss a single (possibly biased) coin \( n \) times, and report the number \( k \) of times it comes up
  \[ \mathcal{K} = \{0, 1, 2, \ldots, n\} \]
  \[ 0 \leq \theta \leq 1 \]
  \[ \text{Bin}(k \mid n, \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k} \]
  \[ \binom{n}{k} = \frac{n!}{(n-k)!k!} \]
Binomial Distributions

- $\theta = 0.250$
- $\theta = 0.500$
- $\theta = 0.900$
Bean Machine (Sir Francis Galton)

http://en.wikipedia.org/wiki/Bean_machine
Categorical Random Variables

- **Multinoulli Distribution**: Single roll of a (possibly biased) die

\[ \mathcal{X} = \{0, 1\}^K, \sum_{k=1}^{K} x_k = 1 \]

\[ \theta = (\theta_1, \theta_2, \ldots, \theta_K), \theta_k \geq 0, \sum_{k=1}^{K} \theta_k = 1 \]

\[ \text{Cat}(x \mid \theta) = \prod_{k=1}^{K} \theta_{x_k} \]

- **Multinomial Distribution**: Roll a single (possibly biased) die \( n \) times, and report the number \( n_k \) of each possible outcome

\[ \text{Mu}(x \mid n, \theta) = \binom{n}{n_1 \ldots n_K} \prod_{k=1}^{K} \theta_{n_k} \]

\[ n_k = \sum_{i=1}^{n} x_{ik} \]
Aligned DNA Sequences
Multinomial Model of DNA