Machine Learning

4771

Instructor: Tony Jebara
Topic 7

• Unsupervised Learning
• Statistical Perspective
• Probability Models
• Discrete & Continuous: Gaussian, Bernoulli, Multinomial
• Maximum Likelihood $\rightarrow$ Logistic Regression
• Conditioning, Marginalizing, Bayes Rule, Expectations
• Classification, Regression, Detection
• Dependence/Independence
• Maximum Likelihood $\rightarrow$ Naïve Bayes
Unsupervised Learning

Classification

Regression, $f(x) = y$

Density/Structure Estimation

Clustering

Feature Selection

Anomaly Detection

Supervised

Unsupervised (can help supervised)
Statistical Perspective

- Several problems with framework so far:
  - Only have input-output approaches (SVM, Neural Net)
  - Pulled non-linear squashing functions out of a hat
  - Pulled loss functions (squared error, etc.) out of a hat
- Better approach for classification?
- What if we have multi-class classification?
- What if other problems, i.e. unobserved values of x,y,etc...
- Also, what if we don’t have a true function?
- Example of Projectile Cannon (c.f. Distal Learning)

- Would like to train a regression function to control a cannon’s angle of fire (y) given target distance (x)
Statistical Perspective

• Example of Projectile Cannon
  (45 degree problem)
  \( x = \text{input target distance} \)
  \( y = \text{output cannon angle} \)

\[
x = \frac{v(0)^2}{g} \sin(2y) + \text{noise}
\]

• What does least squares do?
• Conditional statistical models address this problem...
Probability Models

• Instead of deterministic functions, output is a probability
• Previously: our output was a scalar \( \hat{y} = f(x) = \theta^T x + b \)
• Now: our output is a probability \( p(y) \)
  e.g. a probability bump:

• \( p(y) \) subsumes or is a superset of \( \hat{y} \)
• Why is this representation for our answer more general?
Probability Models

• Instead of deterministic functions, output is a probability
• Previously: our output was a scalar
  \[ \hat{y} = f(x) = \theta^T x + b \]
• Now: our output is a probability
  \[ p(y) \]
e.g. a probability bump:

\[ p(y) \]

• \( p(y) \) subsumes or is a superset of \( \hat{y} \)
• Why is this representation for our answer more general?
  → A deterministic answer \( \hat{y} \) with complete confidence is like putting a probability \( p(y) \) where all the mass is at \( \hat{y} \)!

\[ \hat{y} \Leftrightarrow p(y) = \delta(y - \hat{y}) \]
Probability Models

- Now: our output is a probability density function (pdf) $p(y)$
- Probability Model: a family of pdf’s with adjustable parameters which lets us select one of many $p(y) \rightarrow p(y | \Theta)$
- E.g.: 1-dim Gaussian distribution ‘given’ ‘mean’ parameter $\mu$:
  \[ p(y | \mu) = N(y | \mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu)^2} \]
- Want mean centered on $f(x)$’s value $p(y) = N(y | f(x))$
- Now, linear regression is:
  \[ N(y | f(x)) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-f(x))^2} \]
  \[ = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\Theta^T x - b)^2} \]
Probability Models

• To fit to data, we typically “maximize likelihood” of the probability model

• Log-likelihood = objective function (i.e. negative of cost) for probability models which we want to maximize

• Define (conditional) likelihood as

\[ L(\Theta) = \prod_{i=1}^{N} p(y_i | x_i) \]

or log-Likelihood as

\[ l(\Theta) = \log\left( L(\Theta) \right) = \sum_{i=1}^{N} \log p(y_i | x_i) \]

• For Gaussian \( p(y|x) \), maximum likelihood is least squares!

\[
\begin{align*}
\sum_{i=1}^{N} \log p(y_i | x_i) &= \sum_{i=1}^{N} \log N(y_i | f(x_i)) = \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_i - f(x_i))^2} \\
&= -N \log(\sqrt{2\pi}) - \sum_{i=1}^{N} \frac{1}{2} (y_i - f(x_i))^2
\end{align*}
\]
Probability Models

• Can extend probability model to 2 bumps:
  \[ p(y \mid \Theta) = \frac{1}{2} N(y \mid \mu_1) + \frac{1}{2} N(y \mid \mu_2) \]

• Each mean can be a linear regression fn.
  \[
  p(y \mid x, \Theta) = \frac{1}{2} N(y \mid f_1(x)) + \frac{1}{2} N(y \mid f_2(x)) \\
  = \frac{1}{2} N(y \mid \theta_1^T x + b_1) + \frac{1}{2} N(y \mid \theta_2^T x + b_2)
  \]

• Therefore the (conditional) log-likelihood to maximize is:
  \[
  l(\Theta) = \sum_{i=1}^{N} \log \left( \frac{1}{2} N(y_i \mid \theta_1^T x_i + b_1) + \frac{1}{2} N(y_i \mid \theta_2^T x_i + b_2) \right)
  \]

• Maximize \( l(\theta) \) using gradient ascent
  • Nicely handles the "cannon firing" data
Probability Models

• Now classification: can also go beyond deterministic!
• Previously: wanted output to be binary \( \hat{y} = \{0,1\} \)
• Now: our output is a probability \( p(y) \)
  
  e.g. a probability table:

<table>
<thead>
<tr>
<th>( y = 0 )</th>
<th>( y = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.73</td>
<td>0.27</td>
</tr>
</tbody>
</table>

- This subsumes or is a superset again...
- Consider probability over binary events (coin flips!):
  
  e.g. Bernoulli distribution (i.e. 1x2 probability table) with parameter \( \alpha \)

\[
p(y \mid \alpha) = \alpha^y (1 - \alpha)^{1-y} \quad \alpha \in [0,1]
\]

• Linear classification can be done by setting \( \alpha \) equal to \( f(x) \):

\[
p(y \mid x) = f(x)^y (1 - f(x))^{1-y} \quad f(x) \in [0,1]
\]
Probability Models

- Now linear classification is:
  \[ p(y \mid x) = f(x)^y \left(1 - f(x)\right)^{1-y} \quad f(x) \equiv \alpha \in [0,1] \]

- Log-likelihood is (negative of cost function):
  \[ \sum_{i=1}^{N} \log p(y_i \mid x_i) = \sum_{i=1}^{N} \log f(x_i)^{y_i} \left(1 - f(x_i)\right)^{1-y_i} \]
  \[ = \sum_{i=1}^{N} y_i \log f(x_i) + (1 - y_i) \log \left(1 - f(x_i)\right) \]
  \[ = \sum_{i \in \text{class}_1} \log f(x_i) + \sum_{i \in \text{class}_0} \log \left(1 - f(x_i)\right) \]

- But, need a squashing function since \( f(x) \) in [0,1]

- Use sigmoid or logistic again...
  \[ f(x) = \text{sigmoid} \left( \theta^T x + b \right) \in [0,1] \]

- Called logistic regression \( \rightarrow \) new loss function

- Do gradient descent, similar to logistic output neural net!

- Can also handle multi-layer \( f(x) \) and do backprop again!
Generative Probability Models

• Idea: Extend probability to describe both \( X \) and \( Y \)
• Find probability density function over both: \( p(x, y) \)

E.g. *describe* data with Multi-Dim. Gaussian (later...)

• Called a ‘Generative Model’ because we can use it to synthesize or re-generate data similar to the training data we learned from

• Regression models & classification boundaries are not as flexible
don’t keep info about \( X \)don’t model noise/uncertainty
Properties of PDFs

• Let’s review some basics of probability theory

• First, pdf is a function, multiple inputs, one output:
  \[ p(x_1, \ldots, x_n) \]
  \[ p(x_1 = 0.3, \ldots, x_n = 1) = 0.2 \]

• Function’s output is always non-negative:
  \[ p(x_1, \ldots, x_n) \geq 0 \]

• Can have discrete or continuous or both inputs:
  \[ p(x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 3.1415) \]

• Summing over the domain of all inputs gives unity:
  \[ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) \, dx \, dy = 1 \]
  \[ \sum_{y} \sum_{x} p(x, y) = 1 \]

Continuous \rightarrow \text{integral,} \quad \text{Discrete} \rightarrow \text{sum}
Properties of PDFs

• **Marginalizing**: integrate/sum out a variable leaves a marginal distribution over the remaining ones...
  \[ \sum_y p(x, y) = p(x) \]

• **Conditioning**: if a variable ‘y’ is ‘given’ we get a conditional distribution over the remaining ones...
  \[ p(x \mid y) = \frac{p(x, y)}{p(y)} \]

• **Bayes Rule**: mathematically just redo conditioning but has a deeper meaning (1764)... if we have X being data and θ being a model
  \[ p(\theta \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \theta) p(\theta)}{p(\mathcal{X})} \]
Properties of PDFs

• Expectation: can use pdf p(x) to compute averages and expected values for quantities, denoted by:

\[ E_{p(x)} \{ f(x) \} = \int p(x) f(x) \, dx \quad \text{or} \quad \sum_x p(x) f(x) \]

• Properties:

\[ E \{ cf(x) \} = cE \{ f(x) \} \]
\[ E \{ f(x) + c \} = E \{ f(x) \} + c \]
\[ E \{ E \{ f(x) \} \} = E \{ f(x) \} \]

• Mean: expected value for x

\[ E_{p(x)} \{ x \} = \int_{-\infty}^{\infty} p(x) x \, dx \]

• Variance: expected value of (x-mean)^2, how much x varies

\[ Var \{ x \} = E \{ (x - E \{ x \})^2 \} = E \{ x^2 - 2xE \{ x \} + E \{ x \}^2 \} \]
\[ = E \{ x^2 \} - 2E \{ x \} E \{ x \} + E \{ x \}^2 = E \{ x^2 \} - E \{ x \}^2 \]

**Example:** speeding ticket

<table>
<thead>
<tr>
<th>Fine=0$</th>
<th>Fine=20$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Expected cost of speeding?

f(x=0)=0, f(x=1)=20
p(x=0)=0.8, p(x=1)=0.2


Properties of PDFs

• Covariance: how strongly \( x \) and \( y \) vary together

\[
\text{Cov} \{x, y\} = E \left\{ (x - E \{x\})(y - E \{y\}) \right\} = E \{xy\} - E \{x\} E \{y\}
\]

• Conditional Expectation: \( E \{y \mid x\} = \int y p(y \mid x)\,dy \)

\[
E \left\{ E \{y \mid x\} \right\} = \int_x p(x) \int_y p(y \mid x)\,dy\,dx = E \{y\}
\]

• Sample Expectation: If we don’t have pdf \( p(x,y) \) can approximate expectations using samples of data

\[
E_{p(x)} \left\{ f(x) \right\} \approx \frac{1}{N} \sum_{i=1}^N f(x_i)
\]

• Sample Mean:

\[
E \{x\} \approx \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i
\]

• Sample Var:

\[
E \left\{ (x - E(x))^2 \right\} \approx \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2
\]

• Sample Cov:

\[
E \left\{ (x - E(x))(y - E(y)) \right\} \approx \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})
\]
More Properties of PDFs

• **Independence:** probabilities of independent variables multiply. Denote with the following notation:

\[
\begin{align*}
\text{x} \independent \text{y} & \quad \rightarrow \quad p(x, y) = p(x) p(y) \\
\text{x} \independent \text{y} & \quad \rightarrow \quad p(x \mid y) = p(x)
\end{align*}
\]

also note in this case:

\[
E_{p(x,y)} \{xy\} = \int_x \int_y p(x) p(y) xy \, dx \, dy
\]

\[
= \int_x p(x) x \, dx \int_y p(y) y \, dy = E_{p(x)} \{x\} E_{p(y)} \{y\}
\]

• **Conditional independence:** when two variables become independent only if another is observed

\[
\begin{align*}
\text{x} \independent \text{y} \mid \text{z} & \quad \rightarrow \quad p(x \mid y, z) = p(x \mid z) \\
\text{x} \independent \text{y} \mid \text{z} & \quad \rightarrow \quad p(x \mid y) \neq p(x)
\end{align*}
\]
The IID Assumption

• Most of the time, we will assume that a dataset independent and identically distributed (IID)

• In many real situations, data is generated by some black box phenomenon in an arbitrary order.

• Assume we are given a dataset:
  \[ \mathcal{X} = \{x_1, \ldots, x_N\} \]
  “Independent” means that (given the model \( \Theta \)) the probability of our data multiplies:
  \[ p\left(x_1, \ldots, x_N \mid \Theta \right) = \prod_{i=1}^{N} p_i(x_i \mid \Theta) \]
  “Identically distributed” means that each marginal probability is the same for each data point
  \[ p\left(x_1, \ldots, x_N \mid \Theta \right) = \prod_{i=1}^{N} p_i(x_i \mid \Theta) = \prod_{i=1}^{N} p(x_i \mid \Theta) \]
The IID Assumption

• Bayes rule says likelihood is probability of data given model

\[
p(\theta | \mathbf{x}) = \frac{p(\mathbf{x} | \theta) p(\theta)}{p(\mathbf{x})}
\]

likelihood \hspace{1cm} posterior \hspace{1cm} prior

• The likelihood of \( \mathbf{x} = \{x_1, \ldots, x_N\} \) under IID assumptions is:

\[
p(\mathbf{x} | \Theta) = p(x_1, \ldots, x_N | \Theta) = \prod_{i=1}^{N} p_i(x_i | \Theta) = \prod_{i=1}^{N} p(x_i | \Theta)
\]

• Learn joint distribution \( p(x | \Theta) \) by maximum likelihood:

\[
\Theta^* = \arg\max_{\Theta} \prod_{i=1}^{N} p(x_i | \Theta) = \arg\max_{\Theta} \sum_{i=1}^{N} \log p(x_i | \Theta)
\]

• Learn conditional \( p(y | x, \Theta) \) by max conditional likelihood:

\[
\Theta^* = \arg\max_{\Theta} \prod_{i=1}^{N} p(y_i | x_i, \Theta) = \arg\max_{\Theta} \sum_{i=1}^{N} \log p(y_i | x_i, \Theta)
\]
Uses of PDFs

• **Classification:** have $p(x,y)$ and given $x$. Asked for discrete $y$ output, give most probable one

  $$p(x,y) \rightarrow p(y \mid x) \rightarrow \hat{y} = \arg \max_m p(y = m \mid x)$$

• **Regression:** have $p(x,y)$ and given $x$. Asked for a scalar $y$ output, give most probable or expected one

  $$\left\{ (x_i, y_i) \right\} \rightarrow p(x,y) \rightarrow p(y \mid x)$$

  $$\hat{y} = \left\{ \begin{array}{c} \arg \max_y p(y \mid x) \\ E_{p(y \mid x)} \{y\} \end{array} \right.$$ 

• **Anomaly Detection:** if have $p(x,y)$ and given both $x,y$. Asked if it is similar $\rightarrow$ threshold

  $$p(x,y) \geq \text{threshold} \rightarrow \{ \text{normal}, \text{anomaly} \}$$