Learning and Intelligent Systems

Acting under uncertainty: Bayesian decision theory

Prof. Andreas Krause
Learning and Adaptive Systems (las.ethz.ch)
Acting under Uncertainty

- So far, have seen how we can interpret supervised learning as fitting probabilistic models of the data.
- Today, we’ll see how we can use the estimated models to make decisions.
**Acting under uncertainty**

- Suppose we have estimated a logistic regression model (say, for spam filtering), and obtain $P(Y=\text{spam} \mid X)$

- Further suppose we have three actions: **Spam**, **NotSpam**, and **AskUser**

- Which should we pick?

<table>
<thead>
<tr>
<th>Action</th>
<th>State</th>
<th>Y</th>
<th>N</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>0</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>.5</td>
<td>.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Action</th>
<th>P</th>
<th>Expected cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>0.2</td>
<td>0.2 \cdot 0 + 0.8 \cdot 10 = 8</td>
</tr>
<tr>
<td>N</td>
<td>0.8</td>
<td>0.8 \cdot 0 + 0.2 \cdot 0 = 0.2</td>
</tr>
<tr>
<td>A</td>
<td>0.5</td>
<td>0.5 \cdot 0 + 0.5 \cdot 0 = 0.5</td>
</tr>
</tbody>
</table>

$$p = P(Y=\text{spam} \mid X)$$
Bayesian decision theory

Given:
- Conditional distribution over labels $P(y \mid x)$, $y \in \mathcal{Y}$
- Set of actions $\mathcal{A}$
- Cost function $C : \mathcal{Y} \times \mathcal{A} \rightarrow \mathbb{R}$

Bayesian Decision Theory recommends to pick the action that minimizes the expected cost

$$a^* = \arg \min_{a \in \mathcal{A}} \mathbb{E}_y [C(y, a) \mid x]$$

For discrete case:

$$= \sum_y P(y \mid x) C(y, a)$$

If we had access to the true distribution $P(y \mid x)$, this decision implements the Bayesian optimal decision.

In practice, can only estimate it, e.g., (logistic) regression.
Recall: Logistic regression

Learning:

- Find optimal weights by minimizing logistic loss + regularizer

\[ \hat{w} = \arg \min_w \sum_{i=1}^n \log \left( 1 + \exp(-y_i w^T x_i) \right) + \lambda ||w||_2^2 \]

\[ = \arg \max_w P(w \mid x_1, \ldots, x_n, y_1, \ldots, y_n) \]

Classification:

- Use conditional distribution

\[ P(y \mid x, \hat{w}) = \frac{1}{1 + \exp(-y \hat{w}^T x)} \]
Example: logistic regression

- Est. cond. dist.: \[ P(Y=x \mid x, w) = \text{Ber}(\sigma(w^T x)) \]
- Action set: \[ A = \{+1, -1\} \]
- Cost function: \[ C(y, a) = \begin{cases} 1 & \text{if } y \neq a \\ 0 & \text{otherwise} \end{cases} \]

\[ \mathbb{E}_y \left[ C(Y, a) \mid x \right] = P(Y=1 \mid x, w) \cdot [a = +1] + P(Y=-1 \mid x, w) \cdot [a = +1] \]
\[ = P(Y \neq a \mid x, w) = \frac{1}{1 + \exp(a \cdot w^T x)} \]

\[ a^* = \arg\max_a \mathbb{E}_y \left[ C(Y, a) \mid x \right] = \arg\max_a \left( 1 + \exp(a \cdot w^T x) \right) \]
\[ = \arg\max_a a \cdot w^T x = \text{sign}(w^T x) \]
Optimal decisions for logistic regression

- Est. cond. dist.: \( \hat{P}(y \mid x) = \text{Ber}(y; \sigma(\hat{w}^T x)) \)
- Action set: \( \mathcal{A} = \{+1, -1\} \)
- Cost function: \( C(y, a) = [y \neq a] \)

Then the action that minimizes the expected cost

\[
a^* = \arg \min_{a \in \mathcal{A}} \mathbb{E}_y[C(y, a) \mid x]
\]

is the most likely class:

\[
a^* = \arg \max_y \hat{P}(y \mid x) = \text{sign}(w^T x)
\]
Asymmetric costs

- Est. cond. dist.: $\hat{P}(y \mid x) = \text{Ber}(y; \sigma(\hat{w}^T x))$
- Action set: $\mathcal{A} = \{+1, -1\}$
- Costs:
  
  $C(y, a) = \begin{cases} 
  c_{FP} & \text{if } y = -1 \text{ and } a = +1 \\
  c_{FN} & \text{if } y = +1 \text{ and } a = -1 \\
  0 & \text{otherwise}
  \end{cases}$

  Then the action that minimizes the expected cost is

  $C_+ = \mathbb{E}_y [C(y, +1) \mid x] = p_+ \cdot 0 + (1-p_+) \cdot c_{FP} = (1-p_+) \cdot c_{FP}$

  $C_- = \mathbb{E}_y [C(y, -1) \mid x] = p_+ \cdot c_{FN} + (1-p_+) \cdot 0 = p_+ \cdot c_{FN}$

  
  Predict + $\iff$ $C_+ \leq C_- \iff (1-p_+)c_{FP} \leq p_+ \cdot c_{FN} \iff p_+ \geq \frac{c_{FP}}{c_{FP} + c_{FN}}$
Demo: Asymmetric costs
„Doubtful“ logistic regression

- Est. cond. dist.: \( \hat{P}(y \mid x) = \text{Ber}(y; \sigma(\hat{w}^T x)) \)
- Action set: \( \mathcal{A} = \{ +1, -1, D \} \)
- Cost function: 
  \[
  C(y, a) = \begin{cases} 
    [y \neq a] & \text{if } a \in \{ +1, -1 \} \\
    c & \text{if } a = D 
  \end{cases}
  \]

Then the action that minimizes the expected cost

\[
   a^* = \arg \min_{a \in \mathcal{A}} \mathbb{E}_y[C(y, a) \mid x]
\]

is given by:

\[
   a^* = \begin{cases} 
    y & \text{if } \hat{P}(y \mid x) \geq 1 - c \\
    D & \text{otherwise}
  \end{cases}
\]

- I.e., pick most likely class only if confident enough!
Demo: Doubtful Logistic Regression
Example: linear regression

- Est. cond. dist.: \( P(y|x, \omega) = \mathcal{N}(y; \mathbf{w}^T x, \sigma^2) \)
- Action set: \( A = \mathbb{R} \)
- Cost function: \( C(y, a) = (y - a)^2 \)

\[
\begin{align*}
 g(a) &= \mathbb{E}_y \left[ C(Y, a) \mid x \right] = \mathbb{E}_y \left[ (Y - a)^2 \mid x \right] \\
 \frac{d}{da} g(a) &= \mathbb{E}_y \left[ \frac{d}{da} (Y - a)^2 \mid x \right] \\
 &= \mathbb{E}_y \left[ 2(Y - a) \mid x \right] = 2 \mathbb{E}_y \left[ Y \mid x \right] - 2 \mathbb{E}_y \mathbb{E}_y^a \left[ a \mid x \right] = 0 \\
 \Rightarrow \ a^* &= \mathbb{E} \left[ Y \mid x \right] = \mathbf{w}^T x
\end{align*}
\]
Optimal decisions for LS regression

- Est. cond. dist.: \( \hat{P}(y \mid x) = \mathcal{N}(y; \hat{w}^T x, \sigma^2) \)
- Action set: \( \mathcal{A} = \mathbb{R} \)
- Cost function: \( C(y, a) = (y - a)^2 \)

Then the action that minimizes the expected cost

\[
a^* = \arg \min_{a \in \mathcal{A}} \mathbb{E}_y[C(y, a) \mid x]
\]

is the conditional mean:

\[
a^* = \mathbb{E}_y[y \mid x] = \int \hat{P}(y \mid x) dy = \hat{w}^T x
\]
Example: Asymmetric cost for regression

- Est. cond. dist.: \( \hat{P}(y \mid x) = \mathcal{N}(y; \hat{w}^T x, \sigma^2) \)
- Action set: \( \mathcal{A} = \mathbb{R} \)
- Cost: \( C(y, a) = c_1 \max(y - a, 0) + c_2 \max(a - y, 0) \)

Then the action that minimizes the expected cost

\[
a^* = \arg \min_{a \in \mathcal{A}} \mathbb{E}_y[C(y, a) \mid x]
\]

is:

\[
a^* = \hat{w}^T x + \sigma \cdot \Phi^{-1} \left( \frac{c_1}{c_1 + c_2} \right)
\]

Where \( \Phi(x) = \int_{-\infty}^{x} \mathcal{N}(t; 0, 1) \, dt \)
Demo: Asymmetric cost for regression
Outlook: Active learning

- Labels are expensive (need to ask expert)
- **Want to minimize the number of labels**
Uncertainty sampling

• Simple strategy: Always pick the example that we are most uncertain about

S: we trained model \( P(Y | x) \)

For each unlabeled example \( x_i \), can predict \( P(Y = +1 | x_i) = p_i \)

If \( p_i \approx 0 \) or \( p_i \approx 1 \) \( \Rightarrow \) certain

If \( p_i \approx 0.5 \) \( \Rightarrow \) uncertain

Uncertainty score \( U_i = -1 \times \left| p_i - 0.5 \right| \)

\( \Rightarrow \) obtain label for \( i^* = \text{argmin} \left| w^T x_i \right| \)
Uncertainty sampling

- **Given**: Pool of unlabeled examples \( D_X = \{x_1, \ldots, x_n\} \)
- Also maintain labeled data set \( D \), initially empty
- For \( t=1,2,3,\ldots \)
  - Estimate \( \hat{P}(Y_i \mid x_i) \) given current data \( D \)
  - Pick unlabeled example that we are most uncertain about
    \[ i_t \in \arg\min_i |0.5 - \hat{P}(Y_i \mid x_i)| \]
  - Query label \( y_{i_t} \) and set
    \[ D \leftarrow D \cup \{(x_{i_t}, y_{i_t})\} \]
Demo: Uncertainty sampling
Further comments

- Active learning violates i.i.d. assumption!
- Can get stuck with bad models
- More advanced selection criteria available
  - E.g.: query point that reduces uncertainty of other points as much as possible
- Learn more in Data Mining course
Deriving decision rules

- Bayesian decision theory provides a principled way to derive decision rules from conditional distributions.

- Standard rules arise as special cases:
  - Linear regression: $\hat{w}^T x$
  - Logistic regression: $\text{sign}(\hat{w}^T x)$

- Can accommodate more complex settings:
  - “Doubt” (i.e., requiring sufficient confidence)
  - Asymmetric losses
  - Active learning
  - ...
Summary: Learning through MAP inference

- Start with statistical assumptions on data:
  - Data points modeled as iid (can be relaxed)
- Choose likelihood function
  - **Examples**: Gaussian, student-t, logistic, exponential, ...
  - loss function
- Choose prior
  - **Examples**: Gaussian, Laplace, exponential, ...
  - regularizer
- Optimize for MAP parameters
- Choose hyperparameters (i.e., variance, etc.) through cross-validation
- Make predictions via Bayesian Decision Theory
What you should be able to do

- Understand and apply logistic regression and its variants
- Relate logistic regression and Perceptron/SVM
- Derive MAP estimation problems for different priors and likelihood functions
- Solve resulting optimization problems by applying gradient descent
- Derive decision rules from cost functions via Bayesian decision theory
- Apply uncertainty sampling for binary classification