Introduction to Structured Prediction and Recap of Bayesian

Xintian Han

CDS, NYU

April 3, 2019

Contents

Introduction to Structured Prediction

Multiclass Hypothesis Space: Reframed

- **General [Discrete] Output Space:** $\mathcal{Y}$
- **Base Hypothesis Space:** $\mathcal{H} = \{h : \mathcal{X} \times \mathcal{Y} \to \mathbf{R}\}$
  - $h(x, y)$ gives **compatibility score** between input x and output y
- **Multiclass Hypothesis Space**

$$\mathcal{F} = \left\{ x \mapsto \arg\max_{y \in \mathcal{Y}} h(x, y) \mid h \in \mathcal{H} \right\}$$

- Final prediction function is an $f \in \mathcal{F}$.
- For each $f \in \mathcal{F}$ there is an underlying compatibility score function $h \in \mathcal{H}$.

Part-of-speech (POS) Tagging

- Given a sentence, give a part of speech tag for each word:

x	$\underbrace{\text{[START]}}_{x_0}$	$\underbrace{\text{He}}_{x_1}$	$\underbrace{\text{eats}}_{x_2}$	$\underbrace{\text{apples}}_{x_3}$
y	$\underbrace{\text{[START]}}_{y_0}$	$\underbrace{\text{Pronoun}}_{y_1}$	$\underbrace{\text{Verb}}_{y_2}$	$\underbrace{\text{Noun}}_{y_3}$

- $\mathcal{V} = \{\text{all English words}\} \cup \{\text{[START]}, \text{"."}\}$
- $\mathcal{P} = \{\text{START}, \text{Pronoun}, \text{Verb}, \text{Noun}, \text{Adjective}\}$
- $\mathcal{X} = \mathcal{V}^n$, $n = 1, 2, 3, \ldots$ [Word sequences of any length]
- $\mathcal{Y} = \mathcal{P}^n$, $n = 1, 2, 3, \ldots$[Part of speech sequence of any length]

Structured Prediction

- A **structured prediction** problem is a multiclass problem in which $\mathcal{Y}$ is very large, but has (or we assume it has) a certain structure.
- For POS tagging, $\mathcal{Y}$ grows exponentially in the length of the sentence.
- Typical **structure** assumption: The POS labels form a Markov chain.
  - i.e. $y_{n+1} \mid y_n, y_{n-1}, \ldots, y_0$ is the same as $y_{n+1} \mid y_n$.

Local Feature Functions: Type 1

- A "type 1" **local feature** only depends on
  - the label at a single position, say y_i (label of the ith word) and
  - x at any position
- Example:

$$\begin{aligned}
\phi_1(i, x, y_i) &= 1(x_i = \text{runs})1(y_i = \text{Verb}) \\
\phi_2(i, x, y_i) &= 1(x_i = \text{runs})1(y_i = \text{Noun}) \\
\phi_3(i, x, y_i) &= 1(x_{i-1} = \text{He})1(x_i = \text{runs})1(y_i = \text{Verb})
\end{aligned}$$

Local Feature Functions: Type 2

- A "type 2" **local feature** only depends on
  - the labels at 2 consecutive positions: y_{i-1} and y_i
  - x at any position
- Example:

$$\begin{aligned}\theta_1(i, x, y_{i-1}, y_i) &= 1(y_{i-1} = \text{Pronoun})1(y_i = \text{Verb}) \\ \theta_2(i, x, y_{i-1}, y_i) &= 1(y_{i-1} = \text{Pronoun})1(y_i = \text{Noun})\end{aligned}$$

Local Feature Vector and Compatibility Score

- At each position i in sequence, define the **local feature vector**:

$$\begin{aligned}\Psi_i(x, y_{i-1}, y_i) &= (\phi_1(i, x, y_i), \phi_2(i, x, y_i), \ldots, \\ &\quad \theta_1(i, x, y_{i-1}, y_i), \theta_2(i, x, y_{i-1}, y_i), \ldots)\end{aligned}$$

- **Local compatibility score** for (x, y) at position i is $\langle w, \Psi_i(x, y_{i-1}, y_i) \rangle$.

Sequence Compatibility Score

- The **compatibility score** for the pair of sequences (x, y) is the sum of the local compatibility scores:

$$\begin{aligned} & \sum_i \langle w, \Psi_i(x, y_{i-1}, y_i) \rangle \\ = & \left\langle w, \sum_i \Psi_i(x, y_{i-1}, y_i) \right\rangle \\ = & \langle w, \Psi(x, y) \rangle, \end{aligned}$$

  where we define the sequence feature vector by

$$\Psi(x, y) = \sum_i \Psi_i(x, y_{i-1}, y_i).$$

- So we see this is a special case of linear multiclass prediction.

Sequence Target Loss

- How do we assess the loss for prediction sequence y' for example (x, y)?
- **Hamming loss** is common:

$$\Delta(y, y') = \frac{1}{|y|} \sum_{i=1}^{|y|} 1(y_i \neq y_i')$$

- Could generalize this as

$$\Delta(y, y') = \frac{1}{|y|} \sum_{i=1}^{|y|} \delta(y_i, y_i')$$

What remains to be done?

- To compute predictions, we need to find

$$\arg\max_{y \in \mathcal{Y}} \langle w, \Psi(x, y) \rangle .$$

- This is straightforward for $|\mathcal{Y}|$ small.
- Now $|\mathcal{Y}|$ is exponentially large.
- Because Ψ breaks down into local functions only depending on 2 adjacent labels,
  - we can solve this efficiently using dynamic programming.
  - (Similar to Viterbi decoding.)
- Learning can be done with SGD and a similar dynamic program.

Recap: Bayesian Methods

Bayesian Decision Theory

- Ingredients:
  - **Parameter space** Θ.
  - **Prior**: Distribution $p(\theta)$ on Θ.
  - **Action space** $\mathcal{A}$.
  - **Loss function**: $\ell : \mathcal{A} \times \Theta \to \mathbf{R}$.
- The **posterior risk** of an action $a \in \mathcal{A}$ is

$$\begin{aligned} r(a) &:= \mathbb{E}\left[\ell(\theta, a) \mid \mathcal{D}\right] \\ &= \int \ell(\theta, a) p(\theta \mid \mathcal{D})\, d\theta. \end{aligned}$$

  - It's the **expected loss under the posterior.**
- A **Bayes action** a^* is an action that minimizes posterior risk:

$$r(a^*) = \min_{a \in \mathcal{A}} r(a)$$

The Posterior Predictive Distribution

- Suppose we've already seen data $\mathcal{D}$.
- The **posterior predictive distribution** is given by

$$x \mapsto p(y \mid x, \mathcal{D}) = \int p(y \mid x; \theta) p(\theta \mid \mathcal{D}) \, d\theta.$$

- This is an average of all conditional densities in our family, weighted by the posterior.
- May not have closed form.
- Numerical integral may be hard to compute.

MAP Estimator Versus Posterior Predictive Distribution

- How do we predict by posterior predictive distribution given a new data point x^*?
- We can use $\hat{y} = \arg\max_y p(y \mid x, \mathcal{D})$
- What about our MAP estimator for θ?

$$\hat{\theta} = \arg\max_{\theta} p(\theta \mid \mathcal{D})$$

- We can also predict y by

$$\hat{y} = \arg\max_{y} p(y \mid x; \theta = \hat{\theta})$$

- In general, the predictions from two methods are different.

Questions

Question 1

Question 1. (From DeGroot and Schervish) Let θ denote the proportion of registered voters in a large city who are in favor of a certain proposition. Suppose that the value of θ is unknown, and two statisticians A and B assign to θ the following different prior PDFs $\xi_A(\theta)$ and $\xi_B(\theta)$, respectively:

$$\begin{aligned} \xi_A(\theta) &= 2\theta \quad \text{for } 0 < \theta < 1, \\ \xi_B(\theta) &= 4\theta^3 \quad \text{for } 0 < \theta < 1. \end{aligned}$$

In a random sample of 1000 registered voters from the city, it is found that 710 are in favor of the proposition.

1. Find the posterior distribution that each statistician assigns to θ.
2. Find the Bayes estimate of θ (minimizer of posterior expected loss) for each statistician based on the squared error loss function.
3. Show that after the opinions of the 1000 registered voters in the random sample had been obtained, the Bayes estimates for the two statisticians could not possibly differ by more than 0.002, regardless of the number in the sample who were in favor of the proposition.

Question 1: Solution

Note that both prior distributions are from the Beta family.

1. We have
$$\xi_A(\theta|x) \propto f(x|\theta)\xi_A(\theta) \propto \theta^{711}(1-\theta)^{290}$$
and
$$\xi_B(\theta|x) \propto f(x|\theta)\xi_B(\theta) \propto \theta^{713}(1-\theta)^{290}.$$
Thus the posteriors from A and B are both beta with parameters $(712, 291)$ and $(714, 291)$, respectively.
2. The respective means are $\frac{712}{1003}$ and $\frac{714}{1005}$.
3. In general the two means are given by
$$\frac{a+2}{1003} \quad \text{and} \quad \frac{a+4}{1005}.$$
The difference is less than $2/1000 = .002$.

Question 2 and 3

- **Question 2.** Two statistics students decide to compute 95% confidence intervals for the distribution parameter θ using an i.i.d. sample $X_1, \ldots, X_n$. Student B uses Bayesian methods to find a 95% credible set $[L_B, R_B]$ for θ. Student F uses frequentist methods to find a 95% confidence interval $[L_F, R_F]$ for θ. Both conclude that parameter θ is in their respective intervals with probability at least .95. Who is correct? Explain.
- **Question 3.** Suppose θ has prior distribution Beta(a, b) for some $a, b > 0$. Given θ, suppose we make independent coin flips with heads probability θ. Find values of a, b and the coin flips so that the posterior variance is larger than the prior variance. [Hint: Recall that a Beta(a, b) random variable has variance given by

$$\frac{ab}{(a+b)^2(a+b+1)}.$$

Try $b = 1$.]

Question 2: Solution

- **Question 2.** Two statistics students decide to compute 95% confidence intervals for the distribution parameter θ using an i.i.d. sample $X_1, \ldots, X_n$. Student B uses Bayesian methods to find a 95% credible set $[L_B, R_B]$ for θ. Student F uses frequentist methods to find a 95% confidence interval $[L_F, R_F]$ for θ. Both conclude that parameter θ is in their respective intervals with probability at least .95. Who is correct? Explain.

  Solution:

  - The frequentist student is totally incorrect, since they have misunderstood what a frequentist confidence interval is. Using frequentist methodology, θ is not a random variable, so it doesn't make sense to say it lies in some fixed interval $[L_F, R_F]$. The correct interpretation is that if independent experiments like this were repeated, then at least 95% of the time $[L_F, R_F]$ will contain θ. **That is, the interval is random not θ.**
  - We can say that the Bayesian student is consistent. Recall that to compute the credible set, the Bayesian student had to introduce some prior distribution π on θ. What we can say is if someone believes π is correct, then it is rational, given the data, to conclude that θ will lie in the posterior credible set with probability 95%.

Question 3: Solution

- **Question 3.** Suppose θ has prior distribution Beta(a, b) for some $a, b > 0$. Given θ, suppose we make independent coin flips with heads probability θ. Find values of a, b and the coin flips so that the posterior variance is larger than the prior variance. [Hint: Recall that a Beta(a, b) random variable has variance given by

$$\frac{ab}{(a+b)^2(a+b+1)}.$$

Try $b = 1$.]

Solution: As hinted, let's try $a = 10$, $b = 1$ and 9 coin flips all landing tails. The prior variance is given by

$$\frac{10 \cdot 1}{(10+1)^2(10+1+1)} = \frac{5}{726} \approx .0069$$

while the posterior variance is given by

$$\frac{10 \cdot 10}{(10+10)^2(10+10+1)} = \frac{1}{84} \approx .0119.$$

Question 4

Question 4. What would be the Maximum a Posteriori (MAP) estimator for λ for i.i.d. $\{x_1, x_2, \ldots, x_N\}$ where $x_i \sim \exp(\lambda)$ with prior $\lambda \sim \text{Uniform}[u_0, u_1]$?

Question 4: Solution

- Likelihood: $L(x_1, \ldots, x_N|\lambda) = \lambda^N e^{-\lambda(x_1+\cdots+x_N)}$
- log-likelihood: $\ell(\lambda|x_1, \ldots, x_N) = N \ln \lambda - \lambda(x_1 + \cdots + x_N)$
- $\ell'(\lambda) = \frac{N}{\lambda} - (x_1 + \cdots + x_N) \begin{cases} > 0 & \text{if } 0 < \lambda < 1/\bar{x} = N/(x_1 + \cdots + x_N), \\ = 0 & \text{if } \lambda = 1/\bar{x} \\ < 0 & \text{if } \lambda > 1/\bar{x} \end{cases}$
- Prior: $p(\lambda) = \frac{1}{u_1 - u_0} \mathbb{1}_{[u_0, u_1]}(\lambda)$.
- Posterior: $p(\lambda|x_1, \ldots, x_N) \propto L(x_1, \ldots, x_N|\lambda) p(\lambda) = \lambda e^{-\lambda(x_1+\cdots+x_N)} \mathbb{1}_{[u_0, u_1]}(\lambda)$
- Maximum value of posterior is attained at

$$\lambda = \begin{cases} u_0 & \text{if } u_0 > 1/\bar{x}, \\ 1/\bar{x} & \text{if } u_0 \leqslant 1/\bar{x} \leqslant u_1 \\ u_1 & \text{if } u_1 < 1/\bar{x}. \end{cases}$$