10-601B Recitation 10

Calvin McCarter

November 5, 2015

1 Prediction with HMMs

Suppose we have fitted an HMM to a sequential dataset $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ over T timesteps. The joint distribution over the $\mathbf{X}$ and the hidden states $\mathbf{Q} = \{\mathbf{q}_1, \ldots, \mathbf{q}_T\}$ is then:

$$p(\mathbf{X}, \mathbf{Q}) = p(\mathbf{q}_1)p(\mathbf{x}_1|\mathbf{q}_1)\prod_{t=2}^{T} p(\mathbf{q}_t|\mathbf{q}_{t-1})p(\mathbf{x}|\mathbf{q}_t).$$

The initial probability is given by $p(\mathbf{q}_1)$, the emission probabilities are given by $p(\mathbf{x}_t|\mathbf{q}_t)$, and the transition probabilities are given by $p(\mathbf{q}_t|\mathbf{q}_{t-1})$.

The next state can now be predicted given the current observations by computing $p(\mathbf{x}_{T+1}|\mathbf{X})$. We will see that doing so requires performing the forward pass of the Forward-Backward Algorithm.

$$\begin{aligned}
p(\mathbf{x}_{T+1}|\mathbf{X}) &= \sum_{\mathbf{q}_{T+1}} p(\mathbf{x}_{T+1}, \mathbf{q}_{T+1}|\mathbf{X}) && \text{from } P(A|C) = \sum_B p(A, B|C) \\
&= \sum_{\mathbf{q}_{T+1}} p(\mathbf{x}_{T+1}|\mathbf{q}_{T+1})p(\mathbf{q}_{T+1}|\mathbf{X}) && \text{from } p(\mathbf{x}_{T+1}|\mathbf{q}_{T+1}, \mathbf{X}) = p(\mathbf{x}_{T+1}|\mathbf{q}_{T+1}) \\
&= \sum_{\mathbf{q}_{T+1}} p(\mathbf{x}_{T+1}|\mathbf{q}_{T+1}) \sum_{\mathbf{q}_T} p(\mathbf{q}_{T+1}, \mathbf{q}_T|\mathbf{X})) && \text{from } P(A|C) = \sum_B p(A, B|C) \\
&= \sum_{\mathbf{q}_{T+1}} p(\mathbf{x}_{T+1}|\mathbf{q}_{T+1}) \sum_{\mathbf{q}_T} p(\mathbf{q}_{T+1}|\mathbf{q}_T)p(\mathbf{q}_T|\mathbf{X}) && \text{from } p(\mathbf{q}_{T+1}|\mathbf{q}_T, \mathbf{X}) = p(\mathbf{q}_{T+1}|\mathbf{q}_T) \\
&= \frac{1}{p(\mathbf{X})} \sum_{\mathbf{q}_{T+1}} p(\mathbf{x}_{T+1}|\mathbf{q}_{T+1}) \sum_{\mathbf{q}_T} p(\mathbf{q}_{T+1}|\mathbf{q}_T)p(\mathbf{q}_T, \mathbf{X}) && \text{from Bayes' rule}
\end{aligned}$$

Ignoring the normalization constant $p(\mathbf{X})$, we can now compute $p(\mathbf{x}_{T+1}|\mathbf{X})$ from emission probability $p(\mathbf{x}_{T+1}|\mathbf{q}_{T+1})$, transition probability $p(\mathbf{q}_{T+1}|\mathbf{q}_T)$, and $p(\mathbf{q}_T, \mathbf{X})$. How do compute $p(\mathbf{q}_T, \mathbf{X})$? Notice that the summation over $\mathbf{q}_T$ is shorthand for summing over the probability for different values of $\mathbf{q}_T$, as in $p(\mathbf{q}_T = i, \mathbf{X})$. Each of these probabilities $p(\mathbf{q}_T = i, \mathbf{X})$ is identical to the values $\alpha_T(i)$ from the forward pass of the Forward-Backward algorithm. The ith step of the forward pass computes $p(\mathbf{q}_i, \{\mathbf{x}_1, \ldots, \mathbf{x}_i\})$, so we run it for all T steps, and plug in the results of the Tth step into our formula for $p(\mathbf{x}_{T+1}|\mathbf{X})$.

2 Bias-Variance Tradeoff

Both overfitting and underfitting need to be avoided to minimize generalization error. The tradeoff between the two is illustrated by the bias-variance decomposition. Suppose we are in the regression setting:

$$Y = f(X) + \epsilon, \qquad \mathbb{E}[\epsilon] = 0 \quad \text{Var}[\epsilon] = \sigma^2.$$

Using the training data we compute our estimate $\hat{f}(X)$ and apply it to a test input x. We see that the generalization error when using squared-error loss decomposes into three terms: irreducible error, bias squared, and variance.

$$\begin{aligned}
\text{Error}[x] =& \mathbb{E}\left[(Y - \hat{f}(x))^2 | X = x\right] \\
=& \sigma^2 + \left[\mathbb{E}\,\hat{f}(x) - f(x)\right]^2 + \mathbb{E}\left[\left(\hat{f}(x) - \mathbb{E}\,\hat{f}(x)\right)\right] \\
=& \sigma^2 + \left[\text{Bias}(\hat{f}(x))\right]^2 + \text{Var}\left[\hat{f}(x)\right] \\
=& \text{Irreducible error} + \text{Bias}^2 + \text{Variance}.
\end{aligned}$$

The irreducible error is error from the noise in the true relationship. It cannot be removed even if we recover the true function $f(X)$. The bias generally comes from underfitting. For example, it could come from leaving out useful features or assuming the true model function is linear when it is actually more complex. Variance generally comes from overfitting. The fitted model is actually random due to a finite training dataset; this randomness is especially problematic when we use complex more model classes.

Adapted from *Elements of Statistical Learning* Section 7.3.

I also went through Slides 6-9 of Andrew Ng's Practical Advice for ML.

3 Cross-validation

I went through *Elements of Statistical Learning* Section 7.10.2.