Recitation Notes 10

Konrad Menzel

December 1, 2006

1 Statistical Learning

A statistical learning problem always has the following structure:

- we have a prior on the likelihood of some event B, $\mathbb{P}(B)$
- we then observe that some event A happens, where we know the joint probability of A and B, $\mathbb{P}(A \cap B)$
- we finally update our beliefs about A, where the posterior about B satisfies Bayes' rule

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Whenever we can take the limits, this generalizes to densities of continuous random variables, so that

$$f(y_1|y_2) = \frac{f(y_1, y_2)}{f(y_2)} = \frac{f(y_1, y_2)}{\int_{-\infty}^{\infty} f(y_1, y_2) dy_1}$$

This is in practice often a very nasty problem, because it's typically hard to calculate the integral that gives us the marginal distribution of y_2.

1.1 The Normal Learning Model

A notable exception is the case in which y_1 and y_2 are jointly normal, i.e.

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right)$$

In order to write down the joint density, we first need to invert the covariance matrix,

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \begin{bmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{bmatrix} = \frac{1}{(1-\varrho^2)\sigma_1^2\sigma_2^2} \begin{bmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{bmatrix}$$

where the correlation coefficient ϱ is defined as

$$\varrho := \frac{\sigma_{12}}{\sigma_1\sigma_2}$$

Now before we start working through the algebra, let's for a second go back to Bayes' rule and see where we actually want to get: ideally, we'd like to factor the joint density into

$$f(y_1, y_2) = \tilde{f}(y_1, y_2) g(y_2)$$

where we know how to calculate the integral $\int \tilde{f}(y_1, y_2) dy_1 = \tilde{F}(y_2)$, so that

$$f(y_1|y_2) = \frac{f(y_1, y_2)}{\int_{-\infty}^{\infty} f(y_1, y_2) dy_1} = \frac{\tilde{f}(y_1, y_2) g(y_2)}{\int_{-\infty}^{\infty} \tilde{f}(y_1, y_2) g(y_2) dy_1} = \frac{\tilde{f}(y_1, y_2)}{\tilde{F}(y_2)}$$

Since the conditional density is defined by a ratio, we have to know $f(y_1, y_2)$ only up to a proportionality factor that can depend on everything except y_1.
The density of the bivariate normal is, up to the normalization

$$\begin{aligned} f(y_1, y_2) &\propto \exp\left\{ -\frac{1}{2} \begin{bmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \end{bmatrix}' \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \end{bmatrix} \right\} \\ &= \exp\left\{ 1 - \frac{1}{2(1-\varrho^2)\sigma_1^2\sigma_2^2} \left[\sigma_2^2 (y_1 - \mu_1)^2 - 2\sigma_{12}(y_1 - \mu_1)(y_2 - \mu_2) + \sigma_1^2 (y_2 - \mu_2)^2 \right] \right\} \end{aligned}$$

Picking out the terms inside the exponential which depend on y_1, collecting by powers of y_1, and completing the square, we get

$$\begin{aligned} \sigma_2^2 (y_1 - \mu_1)^2 - 2\sigma_{12}(y_1 - \mu_1)(y_2 - \mu_2) &= \sigma_2^2 y_1^2 - 2[\sigma_2^2 \mu_1 + \sigma_{12}(y_2 - \mu_2)] y_1 + \sigma_2^2 \mu_1^2 + 2\sigma_{12}\mu_1 (y_2 - \mu_2) \\ &= \sigma_2^2 \left[y_1 - \mu_1 - \frac{\sigma_{12}}{\sigma_2^2}(y_2 - \mu_2) \right]^2 + 2\sigma_{12}\mu_1 (y_2 - \mu_2) - \frac{\sigma_{12}^2}{\sigma_2^2}(y_2 - \mu_2)^2 \end{aligned}$$

If we define

$$K(y_2) := \frac{1}{2(1-\varrho^2)\sigma_1^2\sigma_2^2} \left[2\sigma_{12}\mu_1 (y_2 - \mu_2) + \sigma_1^2 (1-\varrho^2)(y_2 - \mu_2)^2 \right] = \frac{\sigma_{12}\mu_1 (y_2 - \mu_2)}{(1-\varrho^2)\sigma_1^2\sigma_2^2} + \frac{(y_2 - \mu_2)^2}{2\sigma_2^2}$$

Now, going back to our expression for the joint density,

$$f(y_1, y_2) \propto \exp\left\{ -\frac{\left[y_1 - \mu_1 - \frac{\sigma_{12}}{\sigma_2^2}(y_2 - \mu_2) \right]^2}{2(1-\varrho^2)\sigma_1^2} - K(y_2) \right\} =: \tilde{f}(y_1, y_2) \exp\{-K(y_2)\}$$

where $\tilde{f}(y_1, y_2)$ looks exactly like a normal density function, except for the normalizing factor.
Now we can finally calculate the density of y_1 conditional on y_2, where we multiply both numerator and denominator by the normalizing factor in order to make the denominator integrate to one:

$$\begin{aligned} f(y_1, y_2) &= \frac{\tilde{f}(y_1, y_2)}{\int \tilde{f}(y_1, y_2) dy_1} = \frac{\frac{\tilde{f}(y_1, y_2)}{\sqrt{2\pi(1-\varrho^2)\sigma_1^2}}}{\int \frac{\tilde{f}(y_1, y_2)}{\sqrt{2\pi(1-\varrho^2)\sigma_1^2}} dy_1} \\ &= \frac{\exp\left\{ -\frac{\left[y_1 - \mu_1 - \frac{\sigma_{12}}{\sigma_2^2}(y_2 - \mu_2) \right]^2}{2(1-\varrho^2)\sigma_1^2} \right\}}{\sqrt{2\pi(1-\varrho^2)\sigma_1^2}} \end{aligned}$$

Now it is easy to see that the density just corresponds to a normal

$$y_1|y_2 \sim \mathcal{N}\left(\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(y_2 - \mu_2), (1-\varrho^2)\sigma_1^2\right)$$

Notice that the update on the expectation of y_1 is the new information in the realization of y_2 times a regression coefficient. This isn't all that surprising since under the assumptions of this model, OLS gives the best linear predictor, and is also Maximum Likelihood under the normality assumption.

1.2 The Jovanovic Example

In the model we saw in the lecture, a worker has productivity $\eta \sim \mathcal{N}\left(M, \frac{1}{H}\right)$, where all parties only observe $y = \eta + \varepsilon$, where $\varepsilon \perp\!\!\!\perp \eta$, and $\varepsilon \overset{iid.}{\sim} \mathcal{N}\left(0, \frac{1}{h}\right)$.
From the conditional distribution formula derived above, we get

$$\begin{aligned}\mathbb{E}[\eta|y] &= \mu_\eta + \frac{\sigma_{\eta y}}{\sigma_y^2}(y - \mu_y) = M + \frac{\sigma_\eta^2}{\sigma_\eta + \sigma_\varepsilon^2}(y - M) \\ &= M + \frac{\frac{1}{H}}{\frac{1}{H} + \frac{1}{h}}(y - M) = M + \frac{h}{H+h}(y - M) = \frac{HM + hy}{H+h}\end{aligned}$$

The conditional expectation of the worker's productivity is therefore simply an average of the marginal expectation and observed output weighted with the respective precision.
The conditional variance from our derivation above simplifies to

$$\mathrm{Var}(\eta|y) = \sigma_\eta^2 - \frac{\sigma_{\eta y}^2}{\sigma_y^2} = \frac{1}{H} - \frac{\frac{1}{H^2}}{\frac{1}{H} + \frac{1}{h}} = \frac{1}{H} - \frac{h}{H(h+H)} = \frac{1}{H+h}$$

Therefore the posterior over the worker's productivity is

$$\eta|y \sim \mathcal{N}\left(\frac{HM + hy}{H+h}, \frac{1}{H+h}\right)$$

We can easily extend this to more than one period. Defining the posteriors recursively, we get that

$$\eta|\{y_1, \ldots, y_{t-1}\} \sim \mathcal{N}\left(M_t, \frac{1}{H_t}\right)$$

where, by induction,

$$\frac{1}{H_t} = \frac{1}{H_{t-1} + h} = \frac{1}{H + (t-1)h}$$

and

$$M_t = \frac{H_{t-1}M_{t-1} + hy_{t-1}}{H_{t-1} + h} = \frac{HM + h\sum_{s=1}^{t-1} y_s}{H + (t-1)h}$$

An important thing to notice is that Bayes' rule ensures that the posterior mean of $\eta|\{y_1, \ldots, y_{t-1}\}$ is a *martingale*,[1] i.e. that its expectation is equal to its prior at $s \leq t$,

$$\mathbb{E}_s M_t = \frac{H_s M_s + h\sum_{r=s}^{t-1} y_r}{H_s + (t-s)h} = \frac{H_s M_s + h(t-s)M_s}{H_s + (t-s)h} = M_s$$

Any reasonable definition of conditional expectations should actually satisfy this martingale property, since otherwise we must have made inefficient use of information at some point.
We can also see that as t increases,

[1] A random process X_t is called a martingale if for every $s < t$, $\mathbb{E}_s[X_t] = X_s$, i.e. there are no predictable shifts in levels

- each additional datapoint carries less weight
- in the limit, the posterior of η doesn't depend on the initial prior anymore, since its weight decreases as more and more information about the worker's productivity comes in.

The following section gives a very prominent model that is a nice illustration of these properties of learning models.

2 The Holmström (1982) Career Concerns Model

In Holmström's (1982) career concerns model, there is a manager who produces

$$y_t = \eta + a_t + \varepsilon_t$$

where $\varepsilon_t \overset{iid.}{\sim} \mathcal{N}\left(0, \frac{1}{h}\right)$ is again an idiosyncratic productivity shock, productivity $\eta \sim N\left(M \frac{1}{H}\right)$ isn't known to anyone in the economy, and in every period the manager can choose an effort level a_t (whith convex costs $c(a_t)$), which is not observed by the market.
The labor market observes past output $\{y_1, \ldots, y_{t-1}\}$ and makes competitive wage offers each period

$$w_t = \mathbb{E}[y_t | y_1, \ldots, y_{t-1}] = \mathbb{E}[\eta | y_1, \ldots, y_{t-1}] + a_t^*$$

where the manager's effort choice a_t^* is common knowledge in equilibrium. Note that the wage is set at the beginning of each period, before the manager takes his decision about effort.
Notably, the current wage w_t doesn't depend on the choice of effort a_t in the same period, so that in the static problem, $a_t^* = 0$. But, and this is the central idea of the model, w_{t+1} does depend on a_t because the manager may be able to manipulate the market's beliefs about his productivity by deviating from the equilibrium effort choice.
From the normal updating formula, we get

$$\begin{aligned} \mathbb{E}[w_{t+1} | y_1, \ldots, y_{t-1}, a_t] &= \frac{H_t M_t + h(\mathbb{E}_t[y_t|\cdot] - M_t - a_t^*)}{H_t + h} \\ &= \frac{H_t M_t + h(a_t - a_t^*)}{H_t + h} =: \lambda_t M_t + (1 - \lambda_t)(a_t - a_t^*) \end{aligned}$$

Since $\lambda_t = \frac{h}{H + (t-1)h}$ depends only on fixed parameters, the choice of effort a_t doesn't depend on past effort choices nor realizations of output. In particular it doesn't matter whether the manager actually knows η beforehand or not.
We assume that the manager maximizes the expected net present value of his lifetime utility over an infinite horizon,

$$\max U := \sum_{s=t}^{\infty} \delta^{s-t} \{\mathbb{E}_t[w_s | \mathbf{a}] - c(a_s)\}$$

Since as noted before, future effort choices don't depend on a_t, the first-order condition with respect to a_t becomes

$$\frac{\partial}{\partial a_t} U = 0 \Leftrightarrow c'(a_t) = \sum_{s=t}^{\infty} \delta^{s-t} \mathbb{E}_t \left[\frac{\partial}{\partial a_t} w_s \middle| \mathbf{a} \right]$$

Therefore

$$c'(a_t) = (1 - \lambda_t) \sum_{s=t}^{\infty} \delta^{s-t} \prod_{r=t}^{s} \lambda_r$$

$$= \frac{h}{H+th} \underbrace{\sum_{s=t}^{\infty} \delta^{s-t} \prod_{r=t}^{s} \frac{H_t}{H_t + rh}}_{=: k_t} \tag{1}$$

where k_t falls to zero at rate $\frac{1}{t}$, so that the marginal cost of effort falls at rate $\frac{1}{t^2}$.
This model has two peculiar features:

- The manager exerts effort even though his current wage doesn't depend on the output in that period. This can be seen as an investment in his productivity signal on the market
- The market's beliefs about the manager's productivity becomes insensitive to current information on output, which decreases the manager's incentive to exert effort over time.

The infinite-horizon model keeps the incentive to invest in the market signal constant (if the manager was going to retire in period T, there would be a steeper decrease in effort), and the decline in effort is solely driven by the decline in the "elasticity" of the posterior with respect to current output.