

Selected Exercises to Bishop Pattern Recognition, 2006 ed.

Benjamin Basseri

2021

1 Chapter 1

1. The goal is to establish that for the optimal weights on an M -degree polynomial regression model using least-squares error, the following identities hold:

$$\sum_{j=0}^M A_{ij} w_j = T_i = \sum_{n=1}^N (x_n)^i t_n$$

$$A_{ij} = \sum_{n=1}^N (x_n)^{i+j}$$

The given error function is quadratic, and hence convex and has a global minimum. Then to find the \mathbf{w} vector that minimizes the error function,

we can take the derivative and set first order conditions. Thus

$$\begin{aligned}
\frac{\partial}{\partial w_i} E(\mathbf{w}) &= \frac{\partial}{\partial w_i} \frac{1}{2} \sum_{n=1}^N (y(x_n, \mathbf{w}) - t_n)^2 \\
&= \frac{\partial}{\partial w_i} \frac{1}{2} \sum_{n=1}^N \left(\sum_{j=0}^M w_j x_n^j - t_n \right)^2 && \text{Expand} \\
&= \sum_{n=1}^N \left(\sum_{j=0}^M w_j x_n^j - t_n \right) x_n^i && \text{Power rule, chain rule} \\
0 &= \sum_{n=1}^N \left(\sum_{j=0}^M x_n^{i+j} w_j - x_n^i t_n \right) && \text{Distribute, set FOC} \\
\sum_{n=1}^N \sum_{j=0}^M x_n^i t_n &= \sum_{n=1}^N \sum_{j=0}^M x_n^{i+j} w_j \\
T_i &= \sum_{j=0}^M A_{ij} w_j
\end{aligned}$$

which is the desired result.

2. The goal is to derive T_i and A_{ij} if the error function uses L2 regularization. Consider the optimum w_i , the i th component of the \mathbf{w} vector. We can again take the derivative of $E(\mathbf{w})$ and set first order conditions, the only difference from before will be that a new term appears in the derivative due to the regularization term.

$$\begin{aligned}
\frac{\partial}{\partial w_i} \left(E(\mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \right) &= \frac{\partial E(\mathbf{w})}{\partial w_i} + \frac{\partial}{\partial w_i} \frac{\lambda}{2} \|\mathbf{w}\|^2 && \text{Linearity of derivative} \\
&= \frac{\partial E(\mathbf{w})}{\partial w_i} + \frac{\partial}{\partial w_i} \frac{\lambda}{2} \sum_{j=0}^M w_j^2 && \text{Re-express norm} \\
&= \frac{\partial E(\mathbf{w})}{\partial w_i} + \lambda w_i
\end{aligned}$$

Applying the result from problem 1, this gives us the equality

$$-\lambda w_i + \sum_{j=0}^M \left(\sum_{n=1}^N x_n^{i+j} \right) w_j = T_i$$

Hence, for each partial derivative of \mathbf{w} , we get a $-\lambda w_i$ on the result of the LHS. If we want to incorporate this into A_{ij} and move the term inside the summation, we must make sure it only applies to the term when $j = i$. So we apply the Kronecker delta δ_{ij} which is 1 when $i = j$ and 0 otherwise:

$$A_{ij} = \sum_{n=1}^N x_n^{i+j} + \delta_{ij}\lambda$$

3. Here we can use the law of total probability to get the marginal probability of selecting apple by conditioning on the box. Let A be the event an apple was selected, and let B be the random variable for which box was selected. Then:

$$P(A) = \sum_{x \in \{r, g, b\}} P(A|B = x)P(B = x)$$

We're given $P(B = r) = 0.2$, $P(B = g) = 0.6$, $P(B = b) = 0.2$. We can also compute the probability of A given each box from its proportion of items in each box which are $3/10$, $1/2$, $3/10$ respectively. This gives us

$$P(A) = \frac{3}{10} \cdot \frac{2}{10} + \frac{3}{10} \cdot \frac{6}{10} + \frac{5}{10} \cdot \frac{2}{10} = \frac{34}{100} = 0.34$$

We're then asked that if an orange was selected, what is the probability it came from the green box. We can solve this using the identity $P(A|B) = P(B|A)P(A)/P(B)$. Let Or be the event an orange is selected:

$$P(B = g|Or) = \frac{P(Or|B = g)P(B = g)}{P(Or)}$$

We know the numerator is $(3/10)(6/10)$ from the given information. All that's left is to compute $P(Or)$:

$$\begin{aligned} P(Or) &= \sum_{x \in \{r, g, b\}} P(Or|B = x)P(B = x) \\ &= \frac{4}{10} \cdot \frac{2}{10} + \frac{3}{10} \cdot \frac{6}{10} + \frac{5}{10} \cdot \frac{2}{10} = \frac{36}{100} \end{aligned}$$

Putting the results together, we have

$$P(B = g|Or) = \frac{\frac{18}{100}}{\frac{36}{100}} = 1/2$$

Interpretation: even though it is slightly more likely to get an orange out of the red box ($4/10$ vs $3/10$) we are much more likely to select the green box ($6/10$ vs $2/10$); these two happen to balance out in this case.

4. Let \hat{x} be the mode of the x distribution, and let \hat{y} be the value such that $g(\hat{y}) = \hat{x}$. If we differentiate p_y , using the product rule we get

$$\begin{aligned}\frac{d}{dy}p_y(y) &= \frac{d}{dy}p_x(g(y))|g'(y)| \\ &= p'_x(g(y))g'(y)|g'(y)| + p_x(g(y)) \cdot \frac{d}{dy}(|g'(y)|)g''(y)\end{aligned}$$

If g is a linear transformation then $g(y) = \lambda y$ and g 's second derivative is 0, which means the second term above is 0. By hypothesis, \hat{x} is a mode of p_x so $p'_x(g(\hat{y})) = 0$ since $g(\hat{y}) = \hat{x}$. Thus, plugging in \hat{y} for y zeros out both terms

$$p'_x(g(\hat{y}))(\dots) + (\dots)g''(\hat{y}) = 0$$

which means that \hat{y} is an extremum for p_y as well.

If g'' is nonzero, i.e. if g is nonlinear, then plugging in \hat{y} still zeros out the first term above but not necessarily the second. So in general \hat{y} need not be a max of p_y if the change of variable is nonlinear.

5. Note that $\mathbb{E}(\mathbb{E}(X)) = \mathbb{E}(X)$. Therefore

$$\begin{aligned}\text{Var}[f(x)] &= \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] \\ &= \mathbb{E}[f(x)^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2] \\ &= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2 \quad \text{Since } \mathbb{E}[f(x)] \text{ is constant} \\ &= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]^2 + \mathbb{E}[f(x)]^2 \\ &= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2\end{aligned}$$

6. Show that if x is independent of y then their covariance is 0.

$$\text{cov}[x, y] = \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] = \mathbb{E}[x]\mathbb{E}[y] - \mathbb{E}[x]\mathbb{E}[y] = 0$$

where $\mathbb{E}_{x,y}[xy] = \mathbb{E}[x]\mathbb{E}[y]$ since x is independent of y .

7. Show that the normalizing constant for $\mathcal{N}(0, \sigma^2)$ is $1/\sqrt{2\pi\sigma^2}$. Consider the integral of the unnormalized density function

$$I = \int_{\mathcal{R}} \exp\left(-\frac{1}{2\sigma^2}x^2\right) dx$$

Not an easy integral. However, we can square it like so:

$$I^2 = \left(\int_{\mathcal{R}} \exp\left(-\frac{1}{2\sigma^2}x^2\right) dx \right) \left(\int_{\mathcal{R}} \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy \right) = \int_{\mathcal{R}^2} \exp\left(-\frac{1}{2\sigma^2}(x^2 + y^2)\right) dx dy$$

So instead of integrating over the real line we integrate over the plane. Now we can switch to polar coordinates, $r^2 = x^2 + y^2$, which will induce a Jacobean of r in the integrand:

$$I^2 = \int_0^{2\pi} \int_0^\infty \exp\left(-\frac{1}{2\sigma^2}r^2\right) r dr d\theta$$

Now we'll do a change of variable:

$$u = \frac{1}{2\sigma^2}r^2, \quad \sigma^2 du = r dr$$

Checking the limits of integration, we have $u \rightarrow 0$ as $r \rightarrow 0$, and $u \rightarrow \infty$ as $r \rightarrow \infty$. So the integral over dr needn't change limits.

$$\begin{aligned} \implies I^2 &= \sigma^2 \int_0^{2\pi} \int_0^\infty e^{-u} du d\theta \\ &= \sigma^2 \int_0^{2\pi} d\theta \int_0^\infty e^{-u} du = \sigma^2 (2\pi) [-e^{-u}]_0^\infty = 2\pi\sigma^2 \end{aligned}$$

Knowing the density function is everywhere positive we can then take the principal square root of I^2 :

$$I = \sqrt{I^2} = \sqrt{2\pi\sigma^2}$$

Hence, dividing by this amount will make the density integrate to 1, giving a valid probability density function.

To show $\mathcal{N}(\mu, \sigma^2)$ is normalized, consider the change of variable $y = x - \mu$. Then $dy = dx$. Hence

$$\begin{aligned} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathcal{R}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) dx &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathcal{R}} \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \sqrt{2\pi\sigma^2} = 1 \end{aligned}$$

8. We want to show the expectation of a normal distribution $\mathbb{E}(\mathcal{N}(\mu, \sigma^2)) = \mu$. By definition we have

$$\mathbb{E}(\mathcal{N}(\mu, \sigma^2)) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^\infty \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) x dx$$

Substitute $y = x - \mu$ to simplify the exponential, then $x = y + \mu$ and we have

$$\mathbb{E}(\mathcal{N}(\mu, \sigma^2)) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}y^2\right)(y + \mu) dy$$

By the linearity of the integral we can distribute across $y + \mu$:

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}y^2\right) y dy \right) + \left(\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}y^2\right) \mu dy \right)$$

I claim the term on the left is 0: the exponential factor in the integrand is even, since it is symmetric about 0, but the y factor is odd. Hence, the y positive interval of the integral cancels out the y negative interval and it integrates to 0.

The integral on the right is simply μ , since μ is constant with respect to dy we can pull it outside the integral, and what's left is the density function for a normal distribution which integrates to 1.

To get the variance of the gaussian, we set

$$\frac{d}{d\sigma^2} \int_{\mathbb{R}} \exp\left(-\frac{1}{2}(x - \mu)^2(\sigma^2)^{-1}\right) dx = \frac{d}{d\sigma^2} \sqrt{2\pi}(\sigma^2)^{1/2}$$

Since all partials of the integrand exist, we can consider the integral a limit as the lower and upper limits approach $\pm\infty$ and apply Leibniz' rule, to push the derivative inside:

$$\begin{aligned} \int_{\mathbb{R}} \frac{\partial}{\partial\sigma^2} \exp\left(-\frac{1}{2}(x - \mu)^2(\sigma^2)^{-1}\right) dx &= \sqrt{2\pi} \left(\frac{1}{2}\right) (\sigma^2)^{-1/2} \\ \int_{\mathbb{R}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) \left(-\frac{1}{2}\right) (x - \mu)^2 (-1) (\sigma^2)^{-2} dx &= \frac{\sqrt{2\pi}}{2\sigma} \end{aligned}$$

Pulling out constants from the integrand gives

$$\left(\frac{1}{\sigma^4}\right) \left(\frac{1}{2}\right) \int_{\mathbb{R}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) (x - \mu)^2 dx = \frac{\sqrt{2\pi}}{2\sigma}$$

We multiply both sides by σ^3 and simplify:

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) (x - \mu)^2 dx = \sigma^2$$

Which is precisely to say that $\mathbb{E}[(x - \mu)^2] = \sigma^2$.

To get the second moment of the gaussian $\mathbb{E}[x^2]$, we can expand the variance:

$$\begin{aligned}\mathbb{E}[(x - \mu)^2] &= \mathbb{E}[x^2 - 2x\mu + \mu^2] \\ &= \mathbb{E}[x^2] - 2\mu\mathbb{E}[x] + \mu^2 \\ &= \mathbb{E}[x^2] - \mu^2 \\ \implies \sigma^2 + \mu^2 &= \mathbb{E}[x^2]\end{aligned}$$

9. Show that the mode of $p \sim \mathcal{N}(\mu, \sigma^2)$ is μ . The mode is precisely where the density reaches its maximum: the argmax of $p(x)$

$$\begin{aligned}x &= 1 \\ \arg \max_x \mathcal{N}(\mu, \sigma^2) &= \arg \max_x \log \mathcal{N}(\mu, \sigma^2) && \text{log is order-preserving} \\ &= \arg \max_x \log \left(\frac{1}{\sqrt{2\pi}} \right) - \frac{1}{2\sigma^2}(x - \mu)^2\end{aligned}$$

Examining the above, we see that whenever $x \neq \mu$, the second term will be nonzero and positive; so subtracting it will lower the value of the expression. Hence the x that maximizes the value is $x = \mu$.

Likewise we can seek the argmax of the MVN:

$$\begin{aligned}\arg \max_x \mathcal{N}(\mu, \Sigma) &= \arg \max_x \log \mathcal{N}(\mu, \Sigma) \\ &= \arg \max_x \log \left(\frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \right) - \frac{1}{2}(x - \mu)^\top \Sigma^{-1} (x - \mu)\end{aligned}$$

Knowing that the covariance matrix Σ and its inverse are positive semidefinite, we can again characterize the density as decreasing in $\|x - \mu\|$. Thus the x that maximizes the value of the function is the one that zeros out the second term, that is when $x = \mu$

10. Show that $\mathbb{E}[x + z] = \mathbb{E}[x] + \mathbb{E}[z]$ if x and z are independent. Of course this is true even if they are dependent by the linearity of expectation.

We can also show this with integrals:

$$\begin{aligned}
\mathbb{E}[x + z] &= \int_{\text{supp } x, z} (x + z) p_{x,z}(x, z) \, dx \, dz \\
&= \int_{\text{supp } x, z} (x + z) p_x(x) p_z(z) \, dx \, dz && \text{using independence} \\
&= \int_{\text{supp } x, z} x p_x(x) p_z(z) \, dx \, dz + \int_{\text{supp } x, z} z p_x(x) p_z(z) \, dx \, dz && \text{linearity of integral} \\
&= \int_{\text{supp } x} \int_{\text{supp } z} x p_x(x) p_z(z) \, dx \, dz + \int_{\text{supp } x} \int_{\text{supp } z} z p_x(x) p_z(z) \, dx \, dz \\
&= \int_{\text{supp } x} x p_x(x) \, dx \int_{\text{supp } z} p_z(z) \, dz + \int_{\text{supp } x} p_x(x) \, dx \int_{\text{supp } z} z p_z(z) \, dz \\
&= \int_{\text{supp } x} x p_x(x) \, dx + \int_{\text{supp } z} z p_z(z) \, dz && \text{pdf's integrate to 1} \\
&= \mathbb{E}[x] + \mathbb{E}[z]
\end{aligned}$$

■