Solutions to Selected Problems In: Pattern Classification by Duda, Hart, Stork

John L. Weatherwax*

February 24, 2008

Problem Solutions

Chapter 2 (Bayesian Decision Theory)

Problem 11 (randomized rules)

Part (a): Let $R(x)$ be the average risk given measurement x. Then since this risk depends on the action and correspondingly the probability of taking such action, this risk can be expressed as

$$R(x) = \sum_{i=1}^{m} R(a_i|x)P(a_i|x)$$

Then the total risk (over all possible x) is given by the average of $R(x)$ or

$$R = \int_{\Omega_x} (\sum_{i=1}^{m} R(a_i|x)P(a_i|x))p(x)dx$$

Part (b): One would pick

$$P(a_i|x) = \begin{cases} 1 & i = \text{ArgMin}_i R(a_i|x) \\ 0 & i = \text{anything else} \end{cases}$$

Thus for every x pick the action that minimizes the local risk. This amounts to a deterministic rule i.e. there is no benefit to using randomness.

*wax@alum.mit.edu

Part (c): Yes, since in a deterministic decision rule setting randomizing some decisions might further improve performance. But how this would be done would be difficult to say.

Problem 13

Let

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ \lambda_r & i = c+1 \\ \lambda_s & \text{otherwise} \end{cases}$$

Then since the definition of risk is expected loss we have

$$R(\alpha_i|x) = \sum_{i=1}^{c} \lambda(a_i|\omega_j) P(\omega_j|x) \qquad i = 1, 2, \ldots, c \tag{1}$$

$$R(\alpha_{c+1}|x) = \lambda_r \tag{2}$$

Now for $i = 1, 2, \ldots, c$ the risk can be simplified as

$$R(\alpha_i|x) = \lambda_s \sum_{j=1,\, j\neq i}^{c} \lambda(a_i|\omega_j) P(\omega_j|x) = \lambda_s (1 - P(\omega_i|x))$$

Our decision function is to choose the action with the smallest risk, this means that we pick the reject option if and only if

$$\lambda_r < \lambda_s (1 - P(\omega_i|x)) \qquad \forall \quad i = 1, 2, \ldots, c$$

This is equivalent to

$$P(\omega_i|x) < 1 - \frac{\lambda_r}{\lambda_s} \qquad \forall \quad i = 1, 2, \ldots, c$$

This can be interpreted as if all posterior probabilities are below a threshold $(1 - \frac{\lambda_r}{\lambda_s})$ we should reject. Otherwise we pick action i such that

$$\lambda_s (1 - P(\omega_i|x)) < \lambda_s (1 - P(\omega_i|x)) \qquad \forall \quad j = 1, 2, \ldots, c,\ j \neq i$$

This is equivalent to

$$P(\omega_i|x) > P(\omega_j|x) \qquad \forall \quad j \neq i$$

This can be interpreted as selecting the class with the largest posteriori probability. In summary we should reject if and only if

$$P(\omega_i|x) < 1 - \frac{\lambda_r}{\lambda_s} \qquad \forall \quad i = 1, 2, \ldots, c \tag{3}$$

Note that if $\lambda_r = 0$ there is no loss for rejecting (equivalent benefit/reward for correctly classifying) and we always reject. If $\lambda_r > \lambda_s$, the loss from rejecting is greater than the loss from missclassifing and we should never reject. This can be seen from the above since $\frac{\lambda_r}{\lambda_s} > 1$ so $1 - \frac{\lambda_r}{\lambda_s} < 0$ so equation 3 will never be satisfied.

Problem 31 (the probability of error)

Part (a): We can assume with out loss of generality that $\mu_1 < \mu_2$. If this is not the case initially change the labels of class one and class two so that it is. We begin by recalling the probability of error P_e

$$\begin{aligned} P_e &= P(\hat{H} = H_1|H_2)P(H_2) + P(\hat{H} = H_2|H_1)P(H_1) \\ &= \left(\int_{-\infty}^{\xi} p(x|H_2)dx\right) P(H_2) + \left(\int_{\xi}^{\infty} p(x|H_1)dx\right) P(H_1)\,. \end{aligned}$$

Here ξ is the decision boundary between the two classes i.e. we classify as class H_1 when $x < \xi$ and classify x as class H_2 when $x > \xi$. Here $P(H_i)$ are the priors for each of the two classes. We begin by finding the Bayes decision boundary ξ under the minimum error criterion. From the discussion in the book, the decision boundary is given by a likelihood ratio test. That is we decide class H_2 if

$$\frac{p(x|H_2)}{p(x|H_1)} \geq \frac{P(H_1)}{P(H_2)}\left(\frac{C_{21} - C_{11}}{C_{12} - C_{22}}\right)\,,$$

and classify the point as a member of H_1 otherwise. The decision boundary ξ is the point at which this inequality is an *equality*. If we use a minimum probability of error criterion the costs C_{ij} are given by $C_{ij} = 1 - \delta_{ij}$, and the decision boundary reduces when $P(H_1) = P(H_2) = \frac{1}{2}$ to

$$p(\xi|H_1) = p(\xi|H_2)\,.$$

If each conditional distributions is Gaussian $p(x|H_i) = \mathcal{N}(\mu_i, \sigma_i^2)$, then the above expression becomes

$$\frac{1}{\sqrt{2\pi}\sigma_1}\exp\{-\frac{1}{2}\frac{(\xi-\mu_1)^2}{\sigma_1^2}\} = \frac{1}{\sqrt{2\pi}\sigma_2}\exp\{-\frac{1}{2}\frac{(\xi-\mu_2)^2}{\sigma_2^2}\}\,.$$

If we assume (for simplicity) that $\sigma_1 = \sigma_2 = \sigma$ the above expression simplifies to $(\xi-\mu_1)^2 = (\xi-\mu_2)^2$, or on taking the square root of both sides of this we get $\xi - \mu_1 = \pm(\xi - \mu_2)$. If we take the plus sign in this equation we get the contradiction that $\mu_1 = \mu_2$. Taking the minus sign and solving for ξ we find our decision threshold given by

$$\xi = \frac{1}{2}(\mu_1 + \mu_2)\,.$$

Again since we have uniform priors ($P(H_1) = P(H_2) = \frac{1}{2}$), we have an error probability given by

$$2P_e = \int_{-\infty}^{\xi}\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2}\frac{(x-\mu_2)^2}{\sigma^2}\}dx + \int_{\xi}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma^2}\}dx\,.$$

or

$$2\sqrt{2\pi}\sigma P_e = \int_{-\infty}^{\xi}\exp\{-\frac{1}{2}\frac{(x-\mu_2)^2}{\sigma^2}\}dx + \int_{\xi}^{\infty}\exp\{-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma^2}\}dx\,.$$

To evaluate the first integral use the substitution $v = \frac{x-\mu_2}{\sqrt{2}\sigma}$ so $dv = \frac{dx}{\sqrt{2}\sigma}$, and in the second use $v = \frac{x-\mu_2}{\sqrt{2}\sigma}$ so $dv = \frac{dx}{\sqrt{2}\sigma}$. This then gives

$$2\sqrt{2\pi}\sigma P_e = \sigma\sqrt{2}\int_{-\infty}^{\frac{\mu_1-\mu_2}{2\sqrt{2}\sigma}} e^{-v^2}dv + \sigma\sqrt{2}\int_{-\frac{\mu_1-\mu_2}{2\sqrt{2}\sigma}}^{\infty} e^{-v^2}dv\,.$$

or

$$P_e = \frac{1}{2\sqrt{\pi}} \int_{-\infty}^{\frac{\mu_1-\mu_2}{2\sqrt{2}\sigma}} e^{-v^2} dv + \frac{1}{2\sqrt{\pi}} \int_{-\frac{\mu_1-\mu_2}{2\sqrt{2}\sigma}}^{\infty} e^{-v^2} dv \,.$$

Now since e^{-v^2} is an even function, we can make the substitution $u = -v$ and transform the second integral into an integral that is the *same* as the first. Doing so we find that

$$P_e = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\frac{\mu_1-\mu_2}{2\sqrt{2}\sigma}} e^{-v^2} dv \,.$$

Doing the same transformation we can convert the first integral into the second and we have the combined representation for P_e of

$$P_e = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\frac{\mu_1-\mu_2}{2\sqrt{2}\sigma}} e^{-v^2} dv = \frac{1}{\sqrt{\pi}} \int_{-\frac{\mu_1-\mu_2}{2\sqrt{2}\sigma}}^{\infty} e^{-v^2} dv \,.$$

Since we have that $\mu_1 < \mu_2$, we see that the lower limit on the second integral is *positive*. In the case when $\mu_1 > \mu_2$ we would switch the names of the two classes and still arrive at the above. In the general case our probability of error is given by

$$P_e = \frac{1}{\sqrt{\pi}} \int_{\frac{a}{\sqrt{2}}}^{\infty} e^{-v^2} dv \,,$$

with a defined as $a = \frac{|\mu_1-\mu_2|}{2\sigma}$. To match exactly the book let $v = \frac{u}{\sqrt{2}}$ so that $dv = \frac{du}{\sqrt{2}}$ and the above integral becomes

$$P_e = \frac{1}{\sqrt{2\pi}} \int_{a}^{\infty} e^{-u^2/2} du \,.$$

In terms of the error function (which is defined as $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$), we have that

$$\begin{aligned} P_e &= \frac{1}{2}\left[\frac{2}{\sqrt{\pi}} \int_{\frac{a}{\sqrt{2}}}^{\infty} e^{-v^2} dv\right] = \frac{1}{2}\left[\frac{2}{\sqrt{\pi}} \int_0^{\infty} e^{-v^2} dv - \frac{2}{\sqrt{\pi}} \int_0^{\frac{a}{\sqrt{2}}} e^{-v^2} dv\right] \\ &= \frac{1}{2}\left[1 - \frac{2}{\sqrt{\pi}} \int_0^{\frac{a}{\sqrt{2}}} e^{-v^2} dv\right] \\ &= \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{|\mu_1-\mu_2|}{2\sqrt{2}\sigma}\right) . \end{aligned}$$

Since the error function is programmed into many mathematical libraries this form is easily evaluated.

Part (b): From the given inequality

$$P_e = \frac{1}{\sqrt{2\pi}} \int_{a}^{\infty} e^{-t^2/2} dt \leq \frac{1}{\sqrt{2\pi}a} e^{-a^2/2} \,,$$

and the definition a as the lower limit of the integration in Part (a) above we have that

$$\begin{aligned} P_e &\leq \frac{1}{\sqrt{2\pi}} \frac{1}{\frac{|\mu_1-\mu_2|}{2\sigma}} e^{-\frac{(\mu_1-\mu_2)^2}{2(4\sigma^2)}} \\ &= \frac{2\sigma}{\sqrt{2\pi}|\mu_1-\mu_2|} \exp\{-\frac{1}{8}\frac{(\mu_1-\mu_2)^2}{\sigma^2}\} \to 0 \,, \end{aligned}$$

as $\frac{|\mu_1-\mu_2|}{\sigma} \to +\infty$. This can happen if μ_1 and μ_2 get far apart of σ shrinks to zero. In each case the classification problem gets easier.

Problem 32 (probability of error in higher dimensions)

As in Problem 31 the classification decision boundary are now the points (x) that satisfy

$$||x-\mu_1||^2 = ||x-\mu_2||^2 .$$

Expanding each side of this expression we find

$$||x||^2 - 2x^t\mu_1 + ||\mu_1||^2 = ||x||^2 - 2x^t\mu_2 + ||\mu_2||^2 .$$

Canceling the common $||x||^2$ from each side and grouping the terms linear in x we have that

$$x^t(\mu_1-\mu_2) = \frac{1}{2}\left(||\mu_1||^2 - ||\mu_2||^2\right) .$$

Which is the equation for a hyperplane.

Problem 34 (error bounds on non-Gaussian data)

Part (a): Since the Bhattacharyya bound is a specialization of the Chernoff bound, we begin by considering the Chernoff bound. Specifically, using the bound $\min(a,b) \le a^\beta b^{1-\beta}$, for $a>0$, $b>0$, and $0 \le \beta \le 1$, one can show that

$$P(\text{error}) \le P(\omega_1)^\beta P(\omega_2)^{1-\beta} e^{-k(\beta)} \quad \text{for} \quad 0 \le \beta \le 1 .$$

Here the expression $k(\beta)$ is given by

$$k(\beta) = \frac{\beta(1-\beta)}{2}(\mu_2-\mu_1)^t\left[\beta\Sigma_1 + (1-\beta)\Sigma_2\right]^{-1}(\mu_2-\mu_1) + \frac{1}{2}\ln\left(\frac{|\beta\Sigma_1+(1-\beta)\Sigma_2|}{|\Sigma_1|^\beta\,|\Sigma_2|^{1-\beta}}\right) .$$

When we desire to compute the Bhattacharyya bound we assign $\beta = 1/2$. For a two class scalar problem with symmetric means and equal variances that is where $\mu_1 = -\mu$, $\mu_2 = +\mu$, and $\sigma_1 = \sigma_2 = \sigma$ the expression for $k(\beta)$ above becomes

$$\begin{aligned} k(\beta) &= \frac{\beta(1-\beta)}{2}(2\mu)\left[\beta\sigma^2 + (1-\beta)\sigma^2\right]^{-1}(2\mu) + \frac{1}{2}\ln\left(\frac{|\beta\sigma^2+(1-\beta)\sigma^2|}{\sigma^{2\beta}\sigma^{2(1-\beta)}}\right) \\ &= \frac{2\beta(1-\beta)\mu^2}{\sigma^2} . \end{aligned}$$

Since the above expression is valid for *all* β in the range $0 \le \beta \le 1$, we can obtain the tightest possible bound by computing the *maximum* of the expression $k(\beta)$ (since the dominant β expression is $e^{-k(\beta)}$. To evaluate this maximum of $k(\beta)$ we take the derivative with respect to β and set that result equal to zero obtaining

$$\frac{d}{d\beta}\left(\frac{2\beta(1-\beta)\mu^2}{\sigma^2}\right) = 0$$

which gives

$$1 - 2\beta = 0$$

or $\beta = 1/2$. Thus in this case the Chernoff bound *equals* the Bhattacharra bound. When $\beta = 1/2$ we have for the following bound on our our error probability

$$P(\text{error}) \leq \sqrt{P(\omega_1)P(\omega_2)}e^{-\frac{\mu^2}{2\sigma^2}} .$$

To further simplify things if we assume that our variance is such that $\sigma^2 = \mu^2$, and that our priors over class are taken to be uniform ($P(\omega_1) = P(\omega_2) = 1/2$) the above becomes

$$P(\text{error}) \leq \frac{1}{2}e^{-1/2} \approx 0.3033 .$$

Chapter 3 (Maximum Likelihood and Bayesian Estimation)

Problem 34

In the problem statement we are told that to operate on a data of "size" n, requires $f(n)$ numeric calculations. Assuming that each calculation requires 10^{-9} seconds of time to compute, we can process a data structure of size n in $10^{-9}f(n)$ seconds. If we want our results finished by a time T then the largest data structure that we can process must satisfy

$$10^{-9}f(n) \leq T \quad \text{or} \quad n \leq f^{-1}(10^9 T)\,.$$

Converting the ranges of times into seconds we find that

$$T = \{1\,\text{sec}, 1\,\text{hour}, 1\,\text{day}, 1\,\text{year}\} = \{1\,\text{sec}, 3600\,\text{sec}, 86400\,\text{sec}, 31536000\,\text{sec}\}\,.$$

Using a very naive algorithm we can compute the largest n such that $f(n)$ by simply incrementing n repeatedly. This is done in the Matlab script `prob_34_chap_3.m` and the results are displayed in Table XXX.

Problem 35 (recursive v.s. direct calculation of means and covariances)

Part (a): To compute $\hat{\mu}_n$ we have to add n, d-dimensional vectors resulting in nd computations. We then have to divide each component by the scalar n, resulting in another n computations. Thus in total we have

$$nd + n = (1+d)n\,,$$

calculations.

To compute the sample covariance matrix, C_n, we have to first compute the differences $x_k - \hat{\mu}_n$ requiring d subtraction for each vector x_k. Since we must do this for each of the n vectors x_k we have a total of nd calculations required to compute all of the $x_k - \hat{\mu}_n$ difference vectors. We then have to compute the outer products $(x_k - \hat{\mu}_n)(x_k - \hat{\mu}_n)'$, which require d^2 calculations to produce each (in a naive implementation). Since we have n such outer products to compute computing them all requires nd^2 operations. We next, have to add all of these n outer products together, resulting in another set of $(n-1)d^2$ operations. Finally dividing by the scalar $n-1$ requires another d^2 operations. Thus in total we have

$$nd + nd^2 + (n-1)d^2 + d^2 = 2nd^2 + nd\,,$$

calculations.

Part (b): We can derive recursive formulas for $\hat{\mu}_n$ and C_n in the following way. First for $\hat{\mu}_n$ we note that

$$\begin{aligned}\hat{\mu}_{n+1} &= \frac{1}{n+1}\sum_{k=1}^{n+1} x_k = \frac{x_{n+1}}{n+1} + \left(\frac{n}{n+1}\right)\frac{1}{n}\sum_{k=1}^{n} x_k \\ &= \frac{1}{n+1}x_{n+1} + \left(\frac{n}{n+1}\right)\hat{\mu}_n\,.\end{aligned}$$

Writing $\frac{n}{n+1}$ as $\frac{n+1-1}{n+1} = 1 - \frac{1}{n+1}$ the above becomes

$$\begin{aligned}\hat{\mu}_{n+1} &= \frac{1}{n+1}x_{n+1} + \hat{\mu}_n - \frac{1}{n+1}\hat{\mu}_n \\ &= \hat{\mu}_n + \left(\frac{1}{n+1}\right)(x_{n+1} - \hat{\mu}_n)\,,\end{aligned}$$

as expected.

To derive the recursive update formula for the covariance matrix C_n we begin by recalling the definition of C_n of

$$C_{n+1} = \frac{1}{n}\sum_{k=1}^{n+1}(x_k - \hat{\mu}_{n+1})(x_k - \hat{\mu}_{n+1})'\,.$$

Now introducing the recursive definition of the mean $\hat{\mu}_{n+1}$ as computed above we have that

$$\begin{aligned}C_{n+1} &= \frac{1}{n}\sum_{k=1}^{n+1}(x_k - \hat{\mu}_n - \frac{1}{n+1}(x_{n+1} - \hat{\mu}_n))(x_k - \hat{\mu}_n - \frac{1}{n+1}(x_{n+1} - \hat{\mu}_n))' \\ &= \frac{1}{n}\sum_{k=1}^{n+1}(x_k - \hat{\mu}_n)(x_k - \hat{\mu}_n)' - \frac{1}{n(n+1)}\sum_{k=1}^{n+1}(x_k - \hat{\mu}_n)(x_{n+1} - \hat{\mu}_n)' \\ &- \frac{1}{n(n+1)}\sum_{k=1}^{n+1}(x_{n+1} - \hat{\mu}_n)(x_k - \hat{\mu}_n)' - \frac{1}{n(n+1)^2}\sum_{k=1}^{n+1}(x_{n+1} - \hat{\mu}_n)(x_{n+1} - \hat{\mu}_n)'\,.\end{aligned}$$

Since $\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$ we see that

$$n\hat{\mu}_n - \frac{1}{n}\sum_{k=1}^{n} x_k = 0 \quad \text{or} \quad \frac{1}{n}\sum_{k=1}^{n}(\hat{\mu}_n - x_k) = 0\,,$$

Using this fact the second and third terms the above can be greatly simplified. For example we have that

$$\sum_{k=1}^{n+1}(x_k - \hat{\mu}_n) = \sum_{k=1}^{n}(x_k - \hat{\mu}_n) + (x_{n+1} - \hat{\mu}_n) = x_{n+1} - \hat{\mu}_n$$

Using this identity and the fact that the fourth term has a summand that does not depend on k we have C_{n+1} given by

$$\begin{aligned}C_{n+1} &= \frac{n-1}{n}\left(\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \hat{\mu}_n)(x_k - \hat{\mu}_n)' + \frac{1}{n-1}(x_{n+1} - \hat{\mu}_n)(x_{n+1} - \hat{\mu}_n)'\right) \\ &- \frac{2}{n(n+1)}(x_{n+1} - \hat{\mu}_n)(x_{n+1} - \hat{\mu}_n) + \frac{1}{n(n+1)}(x_{n+1} - \hat{\mu}_n)(x_{n+1} - \hat{\mu}_n)' \\ &= \left(1 - \frac{1}{n}\right)C_n + \left(\frac{1}{n} - \frac{1}{n(n+1)}\right)(x_{n+1} - \hat{\mu}_n)(x_{n+1} - \hat{\mu}_n)' \\ &= C_n - \frac{1}{n}C_n + \frac{1}{n}\left(\frac{n+1-1}{n+1}\right)(x_{n+1} - \hat{\mu}_n)(x_{n+1} - \hat{\mu}_n)' \\ &= \left(\frac{n-1}{n}\right)C_n + \frac{1}{n+1}(x_{n+1} - \hat{\mu}_n)(x_{n+1} - \hat{\mu}_n)'\,.\end{aligned}$$

which is the result in the book.

Part (c): To compute $\hat{\mu}_n$ using the recurrence formulation we have d subtractions to compute $x_n - \hat{\mu}_{n-1}$ and the scalar division requires another d operations. Finally, addition by the $\hat{\mu}_{n-1}$ requires another d operations giving in total $3d$ operations. To be compared with the $(d+1)n$ operations in the non-recursive case.

To compute the number of operations required to compute C_n recursively we have d computations to compute the difference $x_n - \hat{\mu}_{n-1}$ and then d^2 operations to compute the outer product $(x_n - \hat{\mu}_{n-1})(x_n - \hat{\mu}_{n-1})'$, another d^2 operations to multiply C_{n-1} by $\left(\frac{n-2}{n-1}\right)$ and finally d^2 operations to add this product to the other one. Thus in total we have

$$d + d^2 + d^2 + d^2 + d^2 = 4d^2 + d\,.$$

To be compare with $2nd^2 + nd$ in the non recursive case. For large n the non-recursive calculations are much more computationally intensive.

Chapter 6 (Multilayer Neural Networks)

Problem 39 (derivatives of matrix inner products)

Part (a): We present a slightly different method to prove this here. We desire the gradient (or derivative) of the function

$$\phi \equiv x^T K x \,.$$

The i-th component of this derivative is given by

$$\frac{\partial \phi}{\partial x_i} = \frac{\partial}{\partial x_i}\left(x^T K x\right) = e_i^T K x + x^T K e_i$$

where e_i is the i-th elementary basis function for $\mathbb{R}^n$, i.e. it has a 1 in the i-th position and zeros everywhere else. Now since

$$(e_i^T K x)^T = x^T K^T e_i^T = e_i^T K x \,,$$

the above becomes

$$\frac{\partial \phi}{\partial x_i} = e_i^T K x + e_i^T K^T x = e_i^T \left((K + K^T)x\right) \,.$$

Since multiplying by e_i^T on the left selects the i-th row from the expression to its right we see that the full gradient expression is given by

$$\nabla \phi = (K + K^T)x \,,$$

as requested in the text. Note that this expression can also be proved easily by writing each term in component notation as suggested in the text.

Part (b): If K is symmetric then since $K^T = K$ the above expression simplifies to

$$\nabla \phi = 2Kx \,,$$

as claimed in the book.

Chapter 9 (Algorithm-Independent Machine Learning)

Problem 9 (sums of binomial coefficients)

Part (a): We first recall the binomial theorem which is

$$(x+y)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) x^k y^{n-k} \,.$$

If we let $x = 1$ and $y = 1$ then $x + y = 2$ and the sum above becomes

$$2^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) \,,$$

which is the stated identity. As an aside, note also that if we let $x = -1$ and $y = 1$ then $x + y = 0$ and we obtain another interesting identity involving the binomial coefficients

$$0 = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) (-1)^k \, .$$

Problem 23 (the average of the leave one out means)

We define the leave one out mean $\mu_{(i)}$ as

$$\mu_{(i)} = \frac{1}{n-1} \sum_{j \neq i} x_j \, .$$

This can obviously be computed for every i. Now we want to consider the mean of the leave one out means. To do so first notice that $\mu_{(i)}$ can be written as

$$\begin{aligned} \mu_{(i)} &= \frac{1}{n-1} \sum_{j \neq i} x_j \\ &= \frac{1}{n-1} \left(\sum_{j=1}^{n} x_j - x_i \right) \\ &= \frac{1}{n-1} \left(n \left(\frac{1}{n} \right) \sum_{j=1}^{n} x_j - x_i \right) \\ &= \frac{n}{n-1} \hat{\mu} - \frac{x_i}{n-1} \, . \end{aligned}$$

With this we see that the mean of all of the leave one out means is given by

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \mu_{(i)} &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{n}{n-1} \hat{\mu} - \frac{x_i}{n-1} \right) \\ &= \frac{n}{n-1} \hat{\mu} - \frac{1}{n-1} \left(\frac{1}{n} \right) \sum_{i=1}^{n} x_i \\ &= \frac{n}{n-1} \hat{\mu} - \frac{\hat{\mu}}{n-1} = \hat{\mu} \, , \end{aligned}$$

as expected.

Problem 40 (maximum likelihood estimation with a binomial random variable)

Since k has a binomial distribution with parameters (n', p) we have that

$$P(k) = \left(\begin{array}{c} n' \\ k \end{array} \right) p^k (1-p)^{n'-k} \, .$$

Then the p that maximizes this expression is given by taking the derivative of the above (with respect to p) setting the resulting expression equal to zero and solving for p. We find that this derivative is given by

$$\frac{d}{dp}P(k) = \left(\begin{array}{c} n' \\ k \end{array} \right) kp^{k-1}(1-p)^{n'-k} + \left(\begin{array}{c} n' \\ k \end{array} \right) p^k(1-p)^{n'-k-1}(n'-k)(-1)\,.$$

Which when set equal to zero and solve for p we find that $p = \frac{k}{n'}$, or the empirical counting estimate of the probability of success as we were asked to show.