Machine Learning for Finance – Problem Set 4 Solutions

Neal Parikh

March 13, 2018

Instructions. Do not refer to any outside sources to complete this assignment, in accordance with the honor code. If you discussed any problems with other students, indicate that in your solutions.

1. *Gaussian discriminant analysis.* Consider a dataset $(x_1, y_1), \ldots, (x_N, y_N)$ with $x_i \in \mathbf{R}^n$ and $y_i \in \{0, 1\}$, and consider the following model for the joint distribution $p(x, y)$:

$$\begin{aligned} y &\sim \text{Bernoulli}(\phi) \\ x \,|\, y = 0 &\sim \text{N}(\mu_0, \Sigma) \\ x \,|\, y = 1 &\sim \text{N}(\mu_1, \Sigma), \end{aligned}$$

with parameters ϕ, μ_0, μ_1, and Σ.

(a) Suppose we already have estimates of all the four parameters and now want to make a prediction at a new query point x^{new}. Show that the posterior distribution of the label at x^{new} takes the form of a logistic function and can be written as

$$p(y = 1 \,|\, x) = \frac{1}{1 + \exp(-\theta^T x)},$$

where θ is a function of ϕ, μ_0, μ_1, and Σ. (To get your answer into the form above, you may need to add a constant feature 1 into x_i and consider them as vectors in $\mathbf{R}^{n+1}$.) This implies, for instance, that linear discriminant analysis is a linear classifier.

(b) Show that the maximum likelihood estimates of the model parameters are given by the following expressions:

$$\begin{aligned} \hat{\phi} &= \frac{1}{N} \sum_{i=1}^{N} [y_i = 1], \\ \hat{\mu}_k &= \frac{\sum_{i=1}^{N} [y_i = k] x_i}{\sum_{i=1}^{N} [y_i = k]}, \\ \hat{\Sigma} &= \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_{y_i})(x_i - \mu_{y_i})^T. \end{aligned}$$

Note. You can use that $\nabla f(X) = X^{-1}$ when $f(X) = \log \det X$ with $\mathbf{dom}\, f = \mathbf{S}^n_{++}$, and that $\nabla f(X) = aa^T$ when $f(X) = a^T X a$.

Solution.

(a) By Bayes' rule,

$$p(y = 1 \mid x) = \frac{p(x \mid y = 1)p(y = 1)}{p(x \mid y = 1)p(y = 1) + p(x \mid y = 0)p(y = 0)}.$$

Plugging in

$$p(x \mid y = k) = \frac{1}{(2\pi)^{n/2} \det \Sigma} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma^{-1}(x - \mu_k)\right)$$

and simplifying gives

$$\left(1 + \frac{1 - \phi}{\phi} \exp\left(-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0) + \frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\right)\right)^{-1},$$

which can be rewritten as

$$\left(1 + \exp\left(-\frac{1}{2}\left(-2\mu_0^T \Sigma^{-1} x + \mu_0^T \Sigma^{-1} \mu_0 + 2\mu_1^T \Sigma^{-1} x - \mu_1^T \Sigma^{-1} \mu_1\right) + \log \frac{1 - \phi}{\phi}\right)\right)^{-1}$$

by expanding the quadratic forms and rearranging. This gives that

$$p(y = 1 \mid x) = \frac{1}{1 + \exp(-\theta^T x)},$$

where

$$\theta = \begin{bmatrix} \frac{1}{2}(\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1) - \log \frac{1-\phi}{\phi} \\ \Sigma^{-1} \mu_1 - \Sigma^{-1} \mu_0 \end{bmatrix}.$$

The algebra above is essentially dull symbol pushing, but these final expressions are not arbitrary gibberish, and it is worthwhile to pause and try to interpret them. The first entry corresponds to the intercept, and it has an appealing form involving a term with the difference between the sizes of μ_0 and μ_1 (measured in the Σ^{-1} squared norm, which distorts the geometry of the space in accordance with the skew in the data) and a term involving the log odds ratio of the two classes (which is zero if the classes are equally common and which provides some adjustment to reflect imbalance otherwise).

Similarly, the other parameter is the difference in means of the two classes, transformed by Σ^{-1}. This vector is the normal vector to the hyperplane describing the decision boundary, as given by the sigmoidal expression above for $p(y = 1 \mid x)$.

All the pieces of information we might expect to affect the placement of the decision boundary are reflected here. To make this completely clear, you may want to draw a picture in $\mathbf{R}^2$ capturing this, beginning with the case where $\Sigma^{-1} = I$.

(b) Let $w = (\phi, \mu_0, \mu_1, \Sigma)$. Then the log likelihood is

$$\ell(w) = \sum_{i=1}^{N} \log p(x_i \mid y_i) + \sum_{i=1}^{N} \log p(y_i).$$

Plugging in the mass and density functions for these expressions and dropping constants that do not depend on the parameters gives

$$\sum_{i=1}^{N} \left(\frac{1}{2} \log \frac{1}{\det \Sigma} - \frac{1}{2}(x_i - \mu_{y_i})^T \Sigma^{-1} (x_i - \mu_{y_i}) + y_i \log \phi + (1 - y_i) \log(1 - \phi) \right).$$

To find the maximum likelihood estimates, we must find the gradients of this expression with respect to each of the parameters. The simplest is ϕ, given by

$$\frac{\partial}{\partial \phi} \ell(w) = \sum_{i=1}^{N} \left(\frac{y_i}{\phi} - \frac{1 - y_i}{1 - \phi} \right) = \frac{\sum_i [y_i = 1]}{\phi} - \frac{N - \sum_i [y_i = 1]}{1 - \phi}.$$

For the means, we have

$$\begin{aligned}
\nabla_{\mu_k} \ell(w) &= -\frac{1}{2} \sum_i [y_i = k] \nabla_{\mu_k} (x_i - \mu_k)^T \Sigma^{-1} (x_i - \mu_k) \\
&= -\frac{1}{2} \sum_i [y_i = k] \nabla_{\mu_k} (\mu_k^T \Sigma^{-1} \mu_k - x_i^T \Sigma^{-1} \mu_k - \mu_k^T \Sigma^{-1} x_i) \\
&= -\frac{1}{2} \sum_i [y_i = k] (2\Sigma^{-1} \mu_k - 2\Sigma^{-1} x_i).
\end{aligned}$$

For the last case, we find the gradient with respect to $S = \Sigma^{-1}$ to simplify the derivation:

$$\begin{aligned}
\nabla_S \ell(w) &= \sum_{i=1}^{N} \nabla_S \left(\frac{1}{2} \log \det S - \frac{1}{2}(x_i - \mu_{y_i})^T S (x_i - \mu_{y_i}) \right) \\
&= \sum_{i=1}^{N} \left(\frac{1}{2} S^{-1} - \frac{1}{2}(x_i - \mu_{y_i})(x_i - \mu_{y_i})^T \right) \\
&= \frac{1}{2} \sum_{i=1}^{N} (\Sigma - (x_i - \mu_{y_i})(x_i - \mu_{y_i})^T).
\end{aligned}$$

Setting the expressions above to zero and rearranging gives the desired result.

2. Consider fitting an SVM to a linearly separable training set. Is the SVM guaranteed to choose a decision boundary that separates the positive and negative classes?

Solution. No, the decision boundary may swing to avoid outliers. Specifically, if λ is very small (close to zero), t_i can be large and the constraints will always be satisfied. The optimization problem becomes an unconstrained optimization problem which encourages $\|w\|_2$ to be small regardless of the training error. In that case, the decision boundary does not necessarily separate postive and negative examples.

3. *Logistic regression and SVMs as regularized loss minimization.* It turns out that (regularized) logistic regression and the support vector machine optimize related objective functions.

   (a) Given an example $x \in \mathbf{R}^n$ and a label $y \in \{0, 1\}$, the log likelihood of the example under the logistic regression model is

   $$y \log g(x) + (1 - y) \log(1 - g(x)),$$

   where $g(z) = 1/(1 + \exp(-z))$ is the sigmoid function. Now suppose we want to switch to the convention of using the labels $\tilde{y} \in \{-1, 1\}$. Show that

   $$y \log g(x) + (1 - y) \log(1 - g(x)) = -\log(1 + \exp(-\tilde{y} w^T x)).$$

   In other words, the expression on the righthand side gives the likelihood of a single example when using the $\{-1, 1\}$ label convention.

   (b) Given a training set $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, with $y_i \in \{-1, 1\}$. Show that the maximum likelihood estimate of the parameters is given by minimizing

   $$\sum_{i=1}^{N} \log(1 + \exp(-y_i w^T x_i)).$$

   (c) Consider the function $f(z) = \log(1 + \exp(-z))$. Explain in English what happens to f under the limits $z \to \infty$ and $z \to -\infty$. Sketch or plot the shape of f and indicate asymptotes and intercepts in your sketch.

   (d) Show that the problem

   $$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & y(w^T x + b) \geq 1 - t \\ & t \geq 0 \end{array}$$

   with variables $w \in \mathbf{R}^n$, $b \in \mathbf{R}$, and $t \in \mathbf{R}$ is equivalent to the unconstrained problem

   $$\text{minimize} \quad (1 - y(w^T x + b))_+$$

   with variables w and b.

   (e) Draw the function $f(z) = (1 - z)_+$ and indicate asymptotes and intercepts. The function f is known as *hinge loss*; explain why this makes sense.

   (f) The problem in part (d) was a version of an SVM estimation problem with a single training example. Generalize your derivation for part (d) to derive an unconstrained problem equivalent to the problem

   $$\begin{array}{ll} \text{minimize} & (1/2)\|w\|_2^2 + \lambda \mathbf{1}^T t \\ \text{subject to} & y_i(w^T x_i + b) \geq 1 - t_i, \quad i = 1, \ldots, N \\ & t \succeq 0, \end{array}$$

   with variables $w \in \mathbf{R}^n$, $b \in \mathbf{R}$, $t \in \mathbf{R}^N$.

Solution.

(a) The expression on the lefthand side is

$$y \log g(x) + (1 - y) \log(1 - g(x)).$$

When $y = 1$, this becomes

$$\log(s(w^T x)) = -\log(1 + \exp(-w^T x)).$$

When $y = 0$ (so $\tilde{y} = -1$), it becomes

$$\log(1 - s(w^T x)) = \log\left(\frac{\exp(-w^T x)}{1 + \exp(-w^T x)}\right) = -\log(1 + \exp(w^T x)).$$

Combining the two cases gives the result.

(b) The maximum likelihood estimate of the usual logistic regression model is obtained by maximizing

$$\sum_{i=1}^{N} y_i \log(s(w^T x_i)) + (1 - y_i) \log(1 - s(w^T x_i)),$$

so from the previous part, this is equivalent to minimizing

$$\sum_{i=1}^{N} \log(1 + \exp(-y_i w^T x_i)).$$

(c) The logistic loss $f(z) = \log(1 + \exp(-z))$ goes to 0 as $z \to \infty$ and has the asymptote $-z$ as $z \to -\infty$, *i.e.*,

$$\lim_{z \to -\infty} \frac{f(z)}{z} = -1.$$

(d) We simply apply two problem transformations we have seen previously in reverse. First, we rewrite the margin constraint as

$$t \geq 1 - y(w^T x + b).$$

Recalling the notation $(x)_+ = \max(0, x)$, we can combine this constraint with the constraint $t \geq 0$ via the equivalent constraint

$$t \geq (1 - y(w^T x + b))_+.$$

This is the problem transformation used to formulate a piecewise-linear minimization as an LP, in reverse. (If $t \geq g(x)$ and $t \geq h(x)$, then $t \geq \max(g(x), h(x))$.)

We then apply the epigraph transformation in reverse to eliminate the variable t, giving

$$\text{minimize} \quad (1 - y(w^T x + b))_+,$$

as desired.

(e) Hinge loss is flat (constant at zero) for $z \geq 1$ and then hinges into the downward-sloping line $1 - z$ at $z = 1$. The logistic loss has this same line as an asymptote as the margin goes to $-\infty$.

(f) The equivalent optimization problem is

$$\text{minimize} \quad \sum_{i=1}^{N}(1 - y_i(w^T x_i + b))_+ + (1/2\lambda)\|w\|_2^2.$$

This is simply a regularized loss minimization problem with hinge loss and Tikhonov regularization. Because of the similarities between logistic loss and hinge loss, we should expect Tikhonov-regularized logistic regression to perform quite similarly. They are both solving convex approximations to the nonconvex 0-1 loss minimization problem by using (fairly similar) convex upper bounds to the 0-1 loss as surrogates.

The main difference is analogous to the difference between using ℓ_1 or Tikhonov regularization, in that it has to do with the presence of the sharp kink at $z = 1$ and the relative magnitudes of the loss functions around that point. Specifically, hinge loss, which is obviously nondifferentiable like the ℓ_1 norm, places a greater emphasis than the logistic loss on driving the loss to exactly zero (*i.e.*, driving the margin to 1) than the smooth logistic loss does.

4. *Invariance properties of support vector machines.* Recall that training an SVM involves solving the convex optimization problem

$$\begin{array}{ll} \text{minimize} & (1/2)\|w\|_2^2 + \lambda \mathbf{1}^T t \\ \text{subject to} & y_i(w^T x_i + b) \geq 1 - t_i, \quad i = 1, \ldots, N \\ & t \succeq 0, \end{array} \tag{1}$$

with variables $w \in \mathbf{R}^n$, $b \in \mathbf{R}$, $t \in \mathbf{R}^N$.

Suppose $\theta^\star = (w^\star, b^\star, t^\star)$ is the solution to the problem above. Here, we look at some ways in which the optimization problem we solve to fit an SVM model can be changed without changing the predictions made by the resulting classifier.

(a) *Margin scaling.* One of the decisions that may have seemed arbitrary but possibly significant in the formulation of the SVM was the choice of scaling the margin boundaries to lie at the hyperplanes $w^T x + b = 1$ and $w^T x + b = -1$. Suppose we replace the 1 in the margin constraint of the SVM with some constant $\kappa > 0$ and replace the regularization parameter λ with $\kappa\lambda$, giving the problem

$$\begin{array}{ll} \text{minimize} & (1/2)\|w\|_2^2 + \kappa\lambda \mathbf{1}^T t \\ \text{subject to} & y_i(w^T x_i + b) \geq \kappa - t_i, \quad i = 1, \ldots, N \\ & t \succeq 0. \end{array} \tag{2}$$

Let $\tilde{\theta}^\star = (\tilde{w}^\star, \tilde{b}^\star, \tilde{t}^\star) = \kappa(w^\star, b^\star, t^\star)$.

i. Show that $\tilde{\theta}^\star$ is feasible for (2).

ii. Show that $\tilde{\theta}^\star$ is the solution of (2).

iii. Show that the modified SVM makes the same classification decisions, *i.e.*, that $\mathbf{sign}((\tilde{w}^\star)^T x + \tilde{b}^\star) = \mathbf{sign}((w^\star)^T x + b^\star)$.

Solution.

i. Since $\theta^\star$ is the unique optimal solution of (1), it satisfies that problem's constraints. It is evident from multiplying those inequalities through by $\kappa > 0$ that $\tilde{\theta}^\star$ satisfies the resulting constraints in (2).

ii. This could be shown either via a problem transformation (change of variables) or by a proof by contradiction; here we use the latter. Suppose there exists some solution $\theta' = (w', b', t')$ of (2) achieving better objective value than $\tilde{\theta}^\star$. We can then show that θ'/κ is also feasible for (1), but that

$$\begin{aligned}(1/2)\|w'\|_2^2 + \kappa\lambda \mathbf{1}^T t &< (1/2)\|\kappa w^\star\|_2^2 + \kappa\lambda \mathbf{1}^T(\kappa t^\star) \\ &= \kappa^2((1/2)\|w^\star\|_2^2 + \lambda \mathbf{1}^T t^\star).\end{aligned}$$

Dividing through by κ^2 gives that

$$(1/2)\|w'/\kappa\|_2^2 + \lambda \mathbf{1}^T(t'/\kappa) < (1/2)\|w^\star\|_2^2 + \lambda \mathbf{1}^T t^\star,$$

i.e., that θ' is a better solution to (1) than the actual solution $\theta^\star$, a contradiction. So $\tilde{\theta}^\star$ is optimal for (2).

iii. Observe that

$$\mathbf{sign}((w^\star)^T x + b) = \mathbf{sign}((\kappa w^\star)^T x + \kappa b^\star) = \mathbf{sign}((\tilde{w}^\star)^T x + \tilde{b}^\star),$$

where the second equality follows because $\kappa > 0$.

(b) *Orthogonal invariance of Gaussian kernels.* Particular forms of the SVM, such as the SVM with some choice of kernel, may have additional invariances. Recall the kernelized dual SVM problem

$$\begin{array}{ll}\text{maximize} & \mathbf{1}^T\alpha - (1/2)\sum_{i,j=1}^N y_i y_j \alpha_i \alpha_j K(x_i, x_j) \\ \text{subject to} & \alpha^T y = 0 \\ & 0 \preceq \alpha \preceq \lambda \mathbf{1},\end{array} \tag{3}$$

with variable $\alpha \in \mathbf{R}^N$. Suppose K is chosen to be the Gaussian kernel

$$K(x, z) = \exp\left(-\frac{\|x - z\|_2^2}{2\sigma^2}\right).$$

Let $\mathcal{D} = \{(x_i, y_i) \mid i = 1, \ldots, N\}$ be a training set and let $\tilde{\mathcal{D}} = \{(Qx_i, y_i) \mid i = 1, \ldots, N\}$ be a modified dataset where the input vectors x_i have been transformed by an orthogonal matrix $Q \in \mathbf{R}^{n \times n}$.

i. Let $\alpha^\star$ and $\tilde{\alpha}^\star$ be the solution to the dual problem when using training sets $\mathcal{D}$ and $\tilde{\mathcal{D}}$, respectively. Show that $\alpha^\star = \tilde{\alpha}^\star$.

ii. Let $\hat{f}(x)$ and $\tilde{f}(x)$ denote predictions on example x of SVMs trained on $\mathcal{D}$ and $\tilde{\mathcal{D}}$, respectively. Which of the following are true?

(A) $\hat{f}(x) = \tilde{f}(x)$

(B) $\hat{f}(x) = \tilde{f}(Qx)$

(C) $\hat{f}(Qx) = \tilde{f}(x)$

You can use the following expression for the optimal value of b for $\tilde{f}$:

$$b = -\frac{\max_i[y_i = -1]w^T\varphi(Qx_i) + \min_i[y_i = 1]w^T\varphi(Qx_i)}{2},$$

where φ is the feature map associated with the Gaussian kernel.

Note. The output of φ for a Gaussian kernel is actually infinite dimensional, so the inner product notation x^Tz should be written differently, but this small abuse of notation does not affect the problem.

Solution.

i. Observe that

$$\begin{aligned} K(Qx, Qz) &= \exp((-1/2\sigma^2)\|Qx - Qz\|_2^2) \\ &= \exp((-1/2\sigma^2)\|Q(x - z)\|_2^2) \\ &= \exp((-1/2\sigma^2)(x - z)^TQ^TQ(x - z)) \\ &= \exp((-1/2\sigma^2)(x - z)^T(x - z)) \\ &= \exp((-1/2\sigma^2)\|x - z\|_2^2) \\ &= K(x, z). \end{aligned}$$

Because the dual problem (3) only depends on the training data through these inner products, we have that $\alpha^\star = \tilde{\alpha}^\star$, *i.e.*, the solutions are the same.

ii. Option (B) is correct. Observe that

$$\begin{aligned} w^T\varphi(Qx) &= \sum_{i=1}^N \alpha_i y_i \varphi(Qx_i)^T\varphi(Qx) \\ &= \sum_{i=1}^N \alpha_i y_i K(Qx_i, Qx) \\ &= \sum_{i=1}^N \alpha_i y_i K(x_i, x), \end{aligned}$$

where the final equality follows by the previous part. The given expression for $b^\star$ also simplifies to

$$b^\star = -\frac{1}{2}\left(\max_i[y_i = -1]\sum_{j=1}^N \alpha_j y_j K(x_j, x_i) + \min_i[y_i = 1]\sum_{j=1}^N \alpha_j y_j K(x_j, x_i)\right)$$

using the expression above for $w^T\varphi(Qx)$. From these, it is easy to verify that $\tilde{f}(Qx) = \mathbf{sign}(w^T\varphi(Qx) + b) = \hat{f}(x)$.

This is intuitive: To classify a new point x^{new} using $\tilde{f}$, we have to initially transform it with Q to obtain the same results as $\hat{f}$, the model trained on the raw data.

5. *A simple duality example.* Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & x^2 + 1 \\ \text{subject to} & (x-2)(x-4) \leq 0, \end{array}$$

with variable $x \in \mathbf{R}$.

(a) *Analysis of primal problem.* Give the feasible set, the optimal value, and the optimal solution.

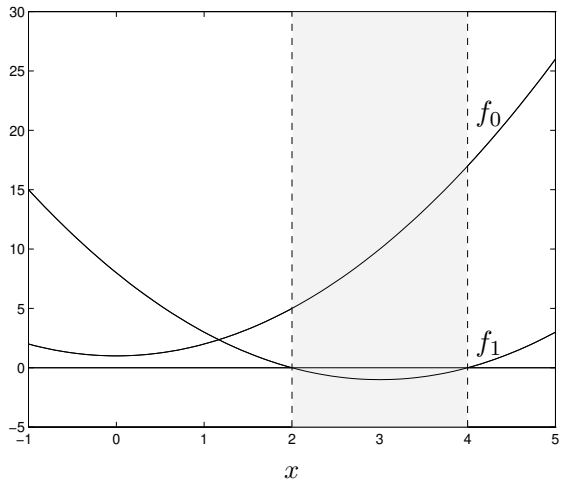
(b) *Lagrangian and dual function.* Plot the objective $x^2 + 1$ versus x. On the same plot, show the feasible set, optimal point and value, and plot the Lagrangian $L(x, \lambda)$ versus x for a few positive values of λ. Verify the lower bound property ($p^\star \geq \inf_x L(x, \lambda)$ for $\lambda \geq 0$). Derive and sketch the Lagrange dual function g.

(c) *Lagrange dual problem.* State the dual problem, and verify that it is a concave maximization problem. Find the dual optimal value and dual optimal solution $\lambda^\star$. Does strong duality hold?

Solution.

(a) The feasible set is the interval $[2, 4]$. The (unique) optimal point is $x^\star = 2$, and the optimal value is $p^\star = 5$.

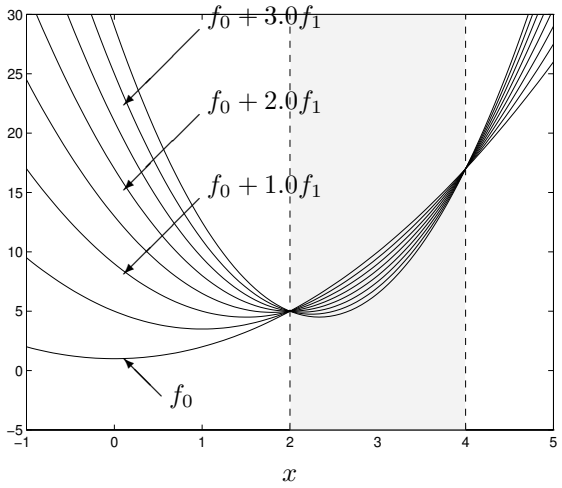
The plot shows f_0 and f_1.



(b) The Lagrangian is

$$L(x, \lambda) = (1 + \lambda)x^2 - 6\lambda x + (1 + 8\lambda).$$

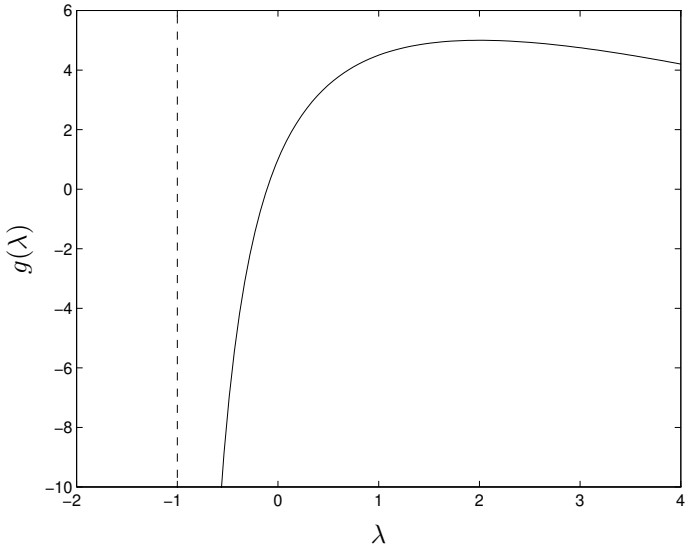
The plot shows the Lagrangian $L(x, \lambda) = f_0 + \lambda f_1$ as a function of x for different values of $\lambda \geq 0$. Note that the minimum value of $L(x, \lambda)$ over x (*i.e.*, $g(\lambda)$) is always less than $p^\star$. It increases as λ varies from 0 toward 2, reaches its maximum at $\lambda = 2$, and then decreases again as λ increases above 2. We have equality $p^\star = g(\lambda)$ for $\lambda = 2$.



For $\lambda > -1$, the Lagrangian reaches its minimum at $\tilde{x} = 3\lambda/(1+\lambda)$. For $\lambda \leq -1$ it is unbounded below. Thus

$$g(\lambda) = \begin{cases} -9\lambda^2/(1+\lambda) + 1 + 8\lambda & \lambda > -1 \\ -\infty & \lambda \leq -1 \end{cases}$$

which is plotted below.



We can verify that the dual function is concave, that its value is equal to $p^\star = 5$ for $\lambda = 2$, and less than $p^\star$ for other values of λ.

(c) The Lagrange dual problem is

$$\begin{array}{ll} \text{maximize} & -9\lambda^2/(1+\lambda) + 1 + 8\lambda \\ \text{subject to} & \lambda \geq 0. \end{array}$$

The dual optimum occurs at $\lambda = 2$, with $d^\star = 5$. So here we can directly observe that strong duality holds (as it must — Slater's constraint qualification is satisfied).

6. *Kernel logistic regression.* Models other than the support vector machine can be kernelized, *i.e.*, put in a form where the kernel trick can be used. Recall that a (regularized) logistic regression model can be fit by solving the convex optimization problem

$$\text{minimize} \quad -\sum_{i=1}^N \log p(y_i \,|\, x_i; w) + (\lambda/2)\|w\|_2^2 \tag{4}$$

with variable $w \in \mathbf{R}^n$, where $p(y = 1 \,|\, x) = s(w^T x)$ and $s(z) = 1/(1+\exp(-z))$ is the sigmoid function. (Here, the inputs x_i are assumed to already contain a constant term and so the separate intercept parameter b is omitted.)

(a) Consider a modified version of logistic regression in which the labels are $\{-1, 1\}$ instead of $\{0, 1\}$, and the conditional probability of the label is given by

$$p(y = k \,|\, x; w) = s(yw^T x)$$

for $k \in \{-1, 1\}$. Show that this implies

$$p(y = 1 \,|\, x) = s(w^T x), \quad p(y = -1 \,|\, x) = 1 - s(w^T x).$$

(b) Show that (4) is equivalent to the problem

$$\begin{array}{ll} \text{minimize} & (\lambda/2)\|w\|_2^2 + \sum_{i=1}^N \log(1 + \exp(-u_i)) \\ \text{subject to} & u_i = y_i w^T x_i, \quad i = 1, \ldots, N, \end{array} \tag{5}$$

with variables $w \in \mathbf{R}^n$, $u \in \mathbf{R}^N$. It should be evident that (5) is convex.

(c) The Lagrangian for (5) is

$$L(w, u, \alpha) = \frac{\lambda}{2}\|w\|_2^2 + \sum_{i=1}^N \log(1 + e^{-u_i}) + \sum_{i=1}^N \alpha_i (u_i - y_i w^T x_i),$$

with dual variables $\alpha \in \mathbf{R}^N$. Find an expression for the $\hat{w}$ minimizing L for a fixed value of α; the expression will be in terms of α. Explain why this expression implies that $0 \preceq \alpha \preceq \mathbf{1}$.

(d) Show that the dual objective is given by

$$g(\alpha) = -\sum_{i=1}^N \left(\alpha_i \log \alpha_i + (1 - \alpha_i) \log(1 - \alpha_i)\right) - \frac{1}{2\lambda} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j x_i^T x_j.$$

(e) Given this dual formulation of the (regularized) logistic regression problem, explain how to kernelize logistic regression. In particular, describe how to train the model and how to efficiently compute $p(y = 1 \,|\, x^{\text{new}})$ for a new test input x^{new}.

Solution.

(a) Clearly, $p(y = 1 \,|\, x) = s(yw^T x) = s(w^T x)$ and $p(y = -1 \,|\, x) = s(yw^T x) = s(-w^T x)$. The last expression can be rewritten

$$s(-w^T x) = \frac{1}{1 + \exp(w^T x)} = \frac{\exp(-w^T x)}{1 + \exp(-w^T x)} = 1 - \frac{1}{1 + \exp(-w^T x)} = 1 - s(w^T x).$$

(b) The problem is evidently convex since it involves minimizing the sum of a quadratic term and a log-sum-exp term subject to linear constraints.

From the previous part, we have

$$\log p(y \mid x) = \log s(yw^T x) = -\log(1 + \exp(-yw^T x)).$$

Introducing new variables $u_i = y_i w^T x_i$ gives the result.

(c) The gradient of the Lagrangian with respect to w and setting to zero gives

$$\nabla_w L(w, u, \alpha) = \lambda w - \sum_{i=1}^N \alpha_i y_i x_i = 0,$$

so

$$\hat{w} = \frac{1}{\lambda} \sum_{i=1}^N \alpha_i y_i x_i. \tag{6}$$

Similarly, setting the partial derivative with respect to u_i to zero gives

$$\frac{\partial}{\partial u_i} L(w, u, \alpha) = -\frac{\exp(-u_i)}{1 + \exp(-u_i)} + \alpha_i = 0,$$

so

$$\hat{u}_i = \log\left(\frac{1}{\alpha_i} - 1\right). \tag{7}$$

Since

$$\alpha_i = \frac{\exp(-\hat{u}_i)}{1 + \exp(-\hat{u}_i)}$$

from above, this implies that $\alpha_i = s(-\hat{u}_i)$, where s is the logistic function, so $0 \preceq \alpha \preceq \mathbf{1}$.

(d) The dual function can be written as

$$\begin{aligned} g(\alpha) &= \inf_{w,u} L(w, u, \alpha) \\ &= (\lambda/2)\|\hat{w}\|_2^2 + \sum_{i=1}^N \log(1 + e^{-\hat{u}_i}) + \sum_{i=1}^N \alpha_i(\hat{u}_i - y_i \hat{w}^T x_i) \\ &= \left((\lambda/2)\|\hat{w}\|_2^2 - \hat{w}^T \sum_{i=1}^N \alpha_i y_i x_i\right) + \left(\sum_{i=1}^N \log(1 + e^{-\hat{u}_i}) + \alpha^T \hat{u}\right). \end{aligned}$$

We now want to eliminate the primal variables w and u to obtain an expression in α. We do this by using (6) and (7).

The variable w can be eliminated from the first grouped term via

$$\begin{aligned}(\lambda/2)\|\hat{w}\|_2^2 - \hat{w}^T \sum_{i=1}^N \alpha_i y_i x_i &= (\lambda/2)\|\hat{w}\|_2^2 - \lambda\|\hat{w}\|_2^2 \\ &= -(\lambda/2)\|\hat{w}\|_2^2 \\ &= -\frac{\lambda}{2}\left\|\frac{1}{\lambda}\sum_{i=1}^N \alpha_i y_i x_i\right\|_2^2 \\ &= -\frac{1}{2\lambda}\sum_{i,j=1}^N \alpha_i\alpha_j y_i y_j x_i^T x_j.\end{aligned}$$

The variable u can be eliminated from the second grouped term via

$$\begin{aligned}&= \sum_{i=1}^N \log(1+e^{-\hat{u}_i}) + \alpha^T\hat{u} \\ &= \sum_{i=1}^N \log\left(1+\exp\left(-\log\left(\frac{1}{\alpha_i}-1\right)\right)\right) + \sum_{i=1}^N \alpha_i \log\left(\frac{1}{\alpha_i}-1\right),\end{aligned}$$

which simplifies down further to

$$-\sum_{i=1}^N (\alpha_i \log\alpha_i + (1-\alpha_i)\log(1-\alpha_i)).$$

with some simple algebra.

Putting these results together gives the dual objective

$$g(\alpha) = -\sum_{i=1}^N (\alpha_i \log\alpha_i + (1-\alpha_i)\log(1-\alpha_i)) - (1/2\lambda)\sum_{i,j=1}^N \alpha_i\alpha_j y_i y_j x_i^T x_j.$$

The full dual problem is then

$$\begin{array}{ll}\text{maximize} & g(\alpha) \\ \text{subject to} & 0 \preceq \alpha \preceq \mathbf{1},\end{array}$$

where the implicit constraints on α have been made explicit.

(e) The dual problem depends on the x_i only via inner products, so the kernel trick can be applied to obtain a kernelized version of logistic regression. Concretely, we solve a modified version of the dual problem given above with $K(x_i, x_j)$ used instead of $x_i^T x_j$.

The prediction for a new test input can be computed via

$$\begin{aligned}p(y=1 \,|\, x^{\text{new}}) &= s(w^T\varphi(x)) \\ &= s\left(\left(\frac{1}{\lambda}\sum_{i=1}^N \alpha_i y_i \varphi(x_i)\right)^T \varphi(x^{\text{new}})\right) \\ &= s\left(\frac{1}{\lambda}\sum_{i=1}^N \alpha_i y_i K(x_i, x^{\text{new}})\right),\end{aligned}$$

where φ is the feature map associated with the kernel K (as before, φ need not be explicit or explicitly evaluated).

7. *Constructing kernels.* Kernelized algorithms let us significantly increase the power of a relatively simple method by allowing it to implicitly work in a high-dimensional space, but a main question is then how to construct kernels for a given problem.

   We have seen two main ways to construct kernels: (a) explicitly define a feature map φ, and (b) use Mercer's theorem. This exercise will build on these by allowing us to construct new kernels from existing ones.

   For each of the functions K below, state whether or not it is a kernel. If so, prove it; otherwise, provide a counterexample.

   (a) $K(x,z) = K_1(x,z) + K_2(x,z)$, where K_1, K_2 are kernels on $\mathbf{R}^n$.

   (b) $K(x,z) = K_1(x,z) - K_2(x,z)$.

   (c) $K(x,z) = \alpha K_1(x,z)$, where $\alpha > 0$.

   (d) $K(x,z) = -\alpha K_1(x,z)$.

   (e) $K(x,z) = K_1(x,z)K_2(x,z)$.

   (f) $K(x,z) = f(x)f(z)$, where $f : \mathbf{R}^n \to \mathbf{R}$. What are the implications if f is a density?

   (g) $K(x,z) = K_3(T(x), T(z))$, where $T : \mathbf{R}^n \to \mathbf{R}^d$.

   (h) $K(x,z) = p(K_1(x,z))$, where p is a polynomial with positive coefficients.

   Solution. All these cases are trivially symmetric because the K_i are symmetric and, in the case of (7f), because multiplication is commutative. By Mercer's theorem, it suffices to show the relevant properties for positive semidefinite matrices G_i, where we will use G_i (for Gram matrix) corresponding to the kernel function K_i.

   (a) Kernel. If $G_1, G_2 \in \mathbf{S}^n_+$, then $G_1 + G_2 \in \mathbf{S}^n_+$.

   (b) Not a kernel. Consider $K_2 = 2K_1$.

   (c) Kernel. If $G_1 \in \mathbf{S}^n_+$, then $\alpha G_1 \in \mathbf{S}^n_+$ when $\alpha > 0$.

   (d) Not a kernel. Consider $\alpha = 1$.

   (e) Kernel. Let $\varphi : \mathbf{R}^n \to \mathbf{R}^d$ be the feature map corresponding to K_1 and $\psi : \mathbf{R}^n \to \mathbf{R}^d$ the feature map for K_2. Then

$$K_1(x,z) = \varphi(x)^T \varphi(z) = \sum_{i=1}^{d} \varphi_i(x)\varphi_i(z).$$

Combining this with a similar expression for K_2 and ψ gives

$$\begin{aligned} K(x,z) &= K_1(x,z)K_2(x,z) \\ &= \left(\sum_{i=1}^{d} \varphi_i(x)\varphi_i(z)\right)\left(\sum_{i=1}^{d} \psi_i(x)\psi_i(z)\right) \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} (\varphi_i(x)\psi_j(x))(\varphi_i(z)\psi_j(z)) \\ &= \sum_{i,j=1}^{d} \tau_{ij}(x)\tau_{ij}(z) \\ &= \tau(x)^T\tau(z), \end{aligned}$$

where $\tau : \mathbf{R}^n \to \mathbf{R}^{d^2}$, the feature map for K, is defined as in the penultimate step.

(f) Kernel. This can be viewed as a one-dimensional feature map.

This seemingly trivial example has interesting consequences: For example, we can take f to be a probability distribution (*i.e.*, a generative model of the raw input data), in which case we are essentially measuring similarity of x and z based on whether a separate model of the data considers them both probable. We can use an arbitrarily complex model of the input data x, which may be an audio signal, a graph of a utility grid, or other structured data, as long as the efficiency of evaluating f (which may require an algorithm in itself) is sufficient for the use case.

(g) Kernel. Since the kernel matrix of K_3 is positive semidefinite for any finite set, it is also positive semidefinite for sets $\{T(x_1), \ldots, T(x_N)\}$ in particular.

(h) Kernel. Combine (7a), (7c), (7e), and (7f) (for the constant term) to see the result.